NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

LOCALLY-WEIGHTED-REGRESSION
SCATTER-PLOT SMOOTHING (LOWESS):
A GRAPHICAL EXPLORATORY
DATA ANALYSIS TECHNIQUE

by

Gary W. Moran

September 1984

Thesis Advisor: P. A. W. Lewis

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Locally-Weighted-Regression Scatter-Plot Smoothing (LOWESS): A Graphical Exploratory Data Analysis Technique

5. TYPE OF REPORT & PERIOD COVERED
Master's Thesis
September 1984

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Gary W. Moran

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Naval Postgraduate School
Monterey, CA 93943

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Naval Postgraduate School
Monterey, CA 93943

12. REPORT DATE
September 1984

13. NUMBER OF PAGES
85

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
curve smoothing, regression, curve fitting

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Statisticians have long used moving average type smoothing and classical regression analysis techniques to reduce the variability in data sets and enhance the visual information presented by scatterplots. This thesis examines the effectiveness of Robuts Locally Weighted Regression Scatterplot Smoothing (LOWESS), a procedure that differs from other techniques because it smooths all of the points and works on unequally as well as equally spaced data. The LOWESS procedure is evaluated by comparing it to

20. (Continued)

previously validated uniform and cosine weighted moving average and least squares regression programs. Interactive APL and FORTRAN programs and detailed user instructions are included for use by interested readers.

A1

Approved for public release; distribution unlimited.

Locally-Weighted-Regression Scatter-Plot Smoothing (LOWESS): a Graphical Exploratory Data Analysis Technique

by

Gary W. Moran
Commander, United States Navy
B.S., United States Naval Academy, 1969

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1984

Author: Gary W Moran
Gary W. Moran

Approved by: Peter A.W. Lewis
P.A.W. Lewis, Thesis Advisor

D.P. Gaver
D.P. Gaver, Second Reader

A.R. Washburn
A.R. Washburn, Chairman,
Department of Operations Research

K.T. Marshall
K.T. Marshall,
Dean of Information and Policy Sciences

ABSTRACT

Statisticians have long used moving average type smoothing and classical regression analysis techniques to reduce the variability in data sets and enhance the visual information presented by scatterplots. This thesis examines the effectiveness of Robust Locally Weighted Regression Scatterplot Smoothing (LOWESS), a procedure that differs from other teqhniques because it smooths all of the points and works on unequally as well as equally spaced data. The LOWESS procedure is evaluated by comparing it to previously validated uniform and cosine weighted moving average and least squares regression programs. Interactive APL and FORTRAN programs and detailed user instructions are included for use by interested readers.

Additional keywords:
Curve smoothing, curve fitting,
APL programming language.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. BACKGROUND

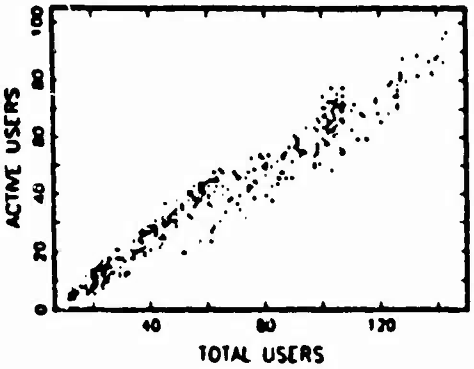
The two dimensional scatter plot has been hailed by many statisticians as being the single most powerful tool used in exploratory data analysis, [Ref. 1]. A scatter plot presents an entire data set in a compact, unambiguous and easily understandable format, in which either:

1. the points lie in a nearly straight line;
2. the points almost lie on a smooth curve;
3. the points are scattered without any apparent correlation between the X variables and the Y variables;
4. the points lie somewhere between (1) or (2) and (3);
5. most of the points lie near a straight line or smooth curve but a few outliers are separated from the rest. [Ref. 2]

These patterns or other hidden peculiarities are much easier to discover during a brief glimpse at a well prepared scatter plot than during an examination of a data table. For example, the strong positive correlation between total users and active users logged on to the W.R. Church computer system, Figure 1.1, is more easily discerned from the plotted points than from the tabulated data[1]. This is a good example of case (1), described above.

Not only does this plot point out the positive trend in the data, it also demonstrates that it is nearly linear and provides a rough estimate of the relationship between the variables.

[1] The table in Figure 1.1 contains only a small portion of the 472 data points included in the plot. A complete listing of the data set takes approximately two pages of text and is not required for demonstration purposes.



TOT	ACT	TOT	ACT
91	59	107	72
92	55	107	67
101	59	115	61
105	63	120	67
104	68	125	76
107	71	123	80
106	72	126	72
106	73	126	73
105	66	126	80
104	70	129	81
104	72	133	83
107	79	138	84
107	76	137	88
105	77	140	90
111	69	142	88
106	71	142	96
106	71	143	98
105	75	139	89

Figure 1.1 Comparison of Data Presentation Methods.

More precise mathematical expressions and confirmatory procedures, including goodness of fit measures, can be obtained by employing classical regression analysis techniques, a logical enhancement of simple scatter plots, Figure 1.2. Numerical quantifications such as the Pearson product moment correlation also provide summaries but can be ambiguous if not accompanied by other information, [Ref. 1 , p 77].

Scatter plots are not invulnerable to misinterpretation. When the scatter of points falls into category (4) or (5), as in Figure 1.3, it may not be possible to judge the true relationship between the variables during a quick glance at the scatter plot, although there obviously is some relationship. Figure 1.3 contains a plot of the first 200 points of test set two (Appendix C) which is used in Chapter III, Section 2 to test LOWESS' ability to follow abrupt changes in curvature.

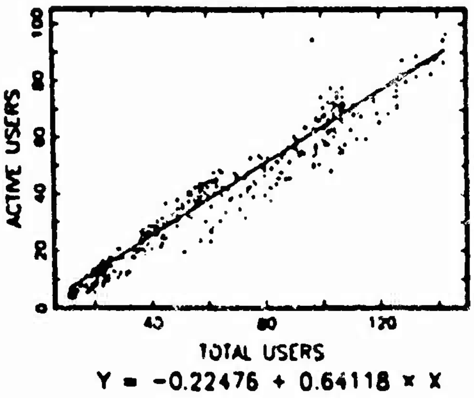


Figure 1.2 Linear Least Squares Regression of Active Users on Total Users Logged on to the W.R. Church Computer System.

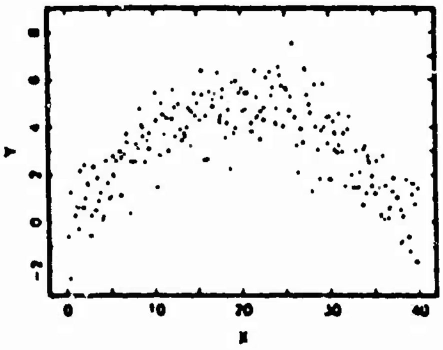
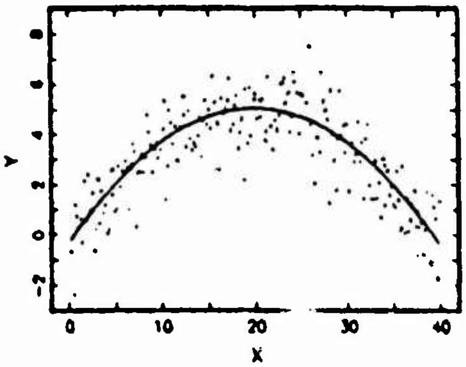


Figure 1.3 Scatter Plot of the First 200 Points of Test Set Two.

Initial inspection of this data suggests the presence of a quadratic type pattern. This impression leads naturally to using the quadratic least squares regression line of Figure 1.4 to describe the dependence of Y on X. The accompanying analysis of variance table lends some support to this choice, since r^2 = .709.

A closer examination of this data reveals, however, that although it looks quadratic, the actual dependence of Y on X



```
Y = +/C × X • 0 1 2   WHERE: C = -0.26565 0.54139 -0 013564

                    ANALYSIS OF VARIANCE TABLE

   SOURCE                    SS          DF          MS          F

GRAND MEAN (SEE NOTE)     2215.056        1
REGRESSION                 523.637        2       261.818    239.551
RESIDUAL                   215.312      197         1.093
TOTAL                     2954.005      200        14.770

THE SIGNIFICANCE LEVEL OF REGRESSION   =  .0000
(SIGNIFICANCE LEVEL = AREA UNDER CURVE BEYOND COMPUTED F)
R SQUARE (SEE NOTE)              =   .709

NOTE: IN WEIGHTED CASE, SEE DESCRIPTION FOR MEANING
```

Figure 1.4   Quadratic Regression on the First 200 Points of Test Set Two.

is not described quite that simply. Figure 1.5 demonstrates this point very clearly. Splitting the data set into three parts at what appear to be logical break points, (x=10,25), and fitting a linear least squares regression line to each, shows that Y is not a single function of X over its entire range. In fact, there appear to be three separate linear trends in this data.

Analyses of this type are seldom undertaken because of the tedium involved in selecting appropriate splitting points once it has been determined that doing so may be helpful.

How then, can an analyst discover the existence of subtle trends or define the shape of unusual patterns contained in a scatter plot? The answer is to use local smoothing procedures rather than global (regression) fitting

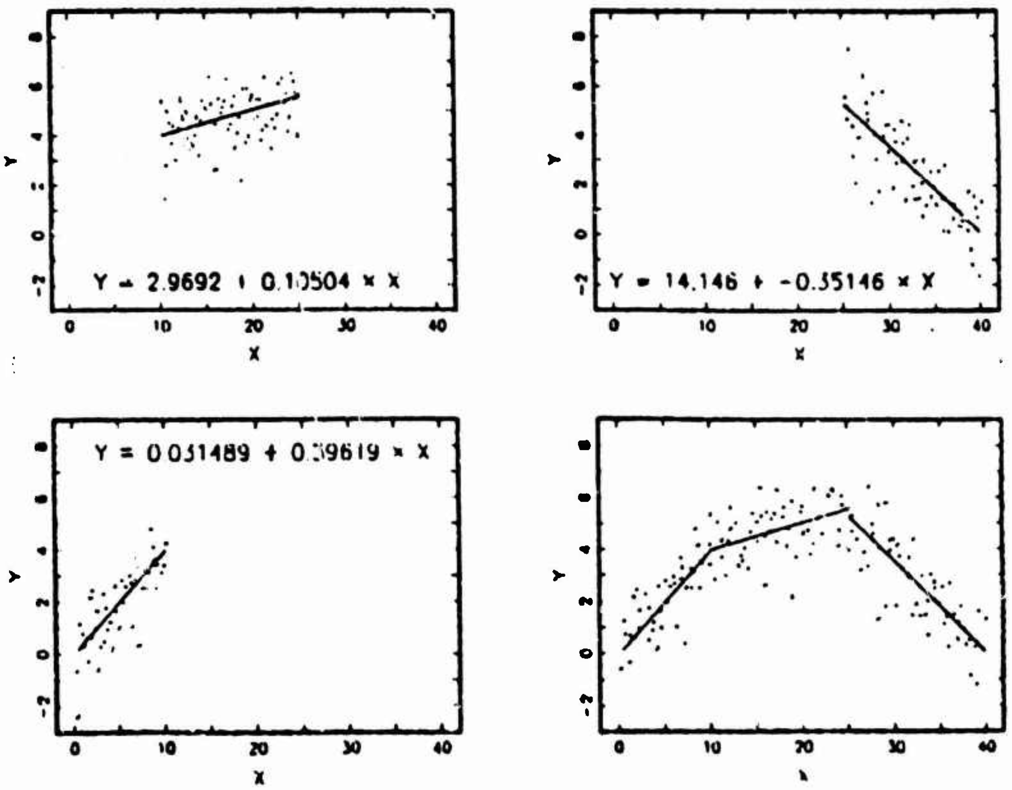


Figure 1.5 Linear Regressions on First 200 Points of Test Set Two Split at X = 10 and 25.

techniques. Using a flexible smoothing procedure that responds to local changes in the data structure allows the data itself to determine the shape of the final curve, as opposed to the classical approach of fitting polynomials which have predetermined shapes.

The Robust Locally Weighted Regression and Scatterplot Smoothing (LOWESS) procedure, [Ref. 3], described in the remainder of this paper, is a very good method for preventing the acceptance of assumptions like the one that led to using the quadratic model in Figure 1.4. The LOWESS smoothing technique applied to this data, the right hand plot of Figure 1.6, shows very clearly, that the dependence of Y on X resembles a combination of three distinct linear functions (the parameter F=.25 will be explained later).

The LOWESS smoothing process has a tendency to round angular corners. The straight lines in the center of each segment suggest linear trends similar to those contained in Figure 1.5.

The major problem with trying to use polynomials to depict subtle trends or to describe unusual relationships in a data set, is that they are neither flexible nor local. By way of example, the points on either extreme of the first of the twc plots in Figure 1.6, have a significant affect on the middle of the fitted polynomials.

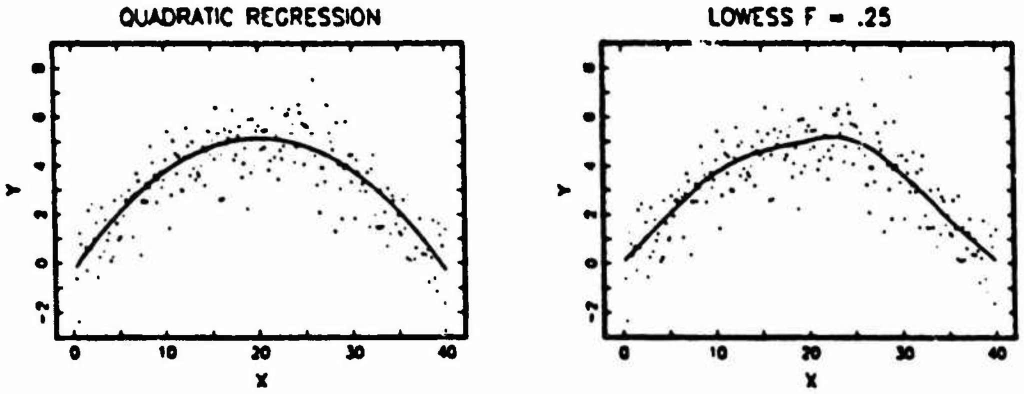


Figure 1.6 Comparison of a Quadratic Regression and LOWESS Smoothing (F = .25) on First 200 Points of Test Set Two.

The LOWESS procedure on the other hand, allows the data points themselves to determine the shape of the smoothed curve. Figure 1.6 also demonstrates that global polynomial regressions have a more difficult time following abrupt pattern changes than do local smoothing procedures.

B. SCOPE

Locally Weighted Regression and Scatterplot Smoothing (LOWESS), introduced by William S. Cleveland in 1977, [Ref. 3], is a generalized extension of the locally fitted polynomial smoothing techniques used for many years in the field of time series[1] analysis.

The essential idea behind the simplest of these classical smoothing techniques is the following. If the data points (Xi,Yi) come from an additive model of the form

$$Y_i = G(X_i) + \epsilon_i$$

where E(ϵi) = 0 and Var(ϵi) = σ^2 and G(Xi) can be approximated locally, over the interval i-m,...i,i+1,...i+m, by the linear function

$$Y_i = B_0(X_i) + B_1(X_i) \times X_i + \epsilon_i$$

then averaging the Yi over this range yields

$$Y_i = \frac{1}{2M+1} \sum_{J=-M}^{M} Y_{i+J}$$

where

$$E(\bar{Y}_i) = B_0(X_i) + B_1(X_i) \times X_i + \bar{\epsilon}_i$$

$$VAR(\bar{Y}_i) = VAR(\epsilon_i) = \frac{\sigma^2}{2M+1}$$

If the assumption that the ϵi are uncorrelated is true, then this moving average process produces estimated $\hat{Y}$i's that are unbiased and have smaller variance than the raw Yi's. This technique makes it easier to distinguish G(Xi) through the noise (ϵi). Using a bandwidth, M, larger than the interval

[1] A time series is a sequence of random variables Yi which are naturally ordered by time (i) and can therefore be presented as a scatter plot of Yi versus i. Although i is usually the integers, missing values can occur.

over which the linearity assumption holds, will introduce bias into the results. [Ref. 4]

The purpose of this thesis is to translate the generalization of classical smooting techniques proposed by Cleveland [Ref. 3], and expounded upon by Chambers et al [Ref. 1], into user friendly computer programs available for use as exploratory data analysis tools by students and faculty of the Naval Postgraduate School.

LOWESS, written in APL, an acronym for "A PROGRAMMING LANGUAGE," was designed to be used alone or in conjunction with the IBM GRAFSTAT statistical graphics package. GRAFSTAT, an experimental program, currently under development by the IBM Watson Reaearch Center, is available at the Naval Postgraduate School for test and evaluation purposes [Ref. 5]. All graphs contained in this paper were produced by the GENERAL PLOT function of the GRAFSTAT program.

LOWS, a modification of LOWESS, when used in conjunction with GRAFSTAT and expanded versions of the DRAFTSMAN DISPLAY programs described in [Ref. 6], enhances an already powerful exploratory data analysis package.

A FORTRAN version of the basic LOWESS program was designed to be used in conjunction with either DISPLA [Ref. 7], or any other W.R. Church computer system supported graphing package.

These programs are interactive and can be used easily by individuals who have little or no APL or FORTRAN programming skills. Users who are well versed in these languages should be able to modify them to provide tailor made outputs, expand their capabilities or incorporate them into other analysis packages.

Detailed user instructions are contained in Chapters IV and V while examples of their use are presented in Chapter III. Users who are interested in the mathematical details of Robust Locally Weighted Regression and Scatterplot Smoothing should read Chapter II.

II. TECHNICAL DESCRIPTION OF LOWESS

A. OVERVIEW

Locally Weighted Regression Scatterplot Smoothing (LOWESS), is a generalized extension of locally fitted polynomial smoothing techniques used by many statisticians in time series analysis [1]. Unlike its predecessors, however, LOWESS was designed to work on unequally as well as equally spaced X's. It also contains a robust fitting procedure that guards against possible distortion of the smoothed curve by outlier points. The general procedure used by Cleveland is an adaptation of iterated least squares regression techniques developed by Albert Beaton and John Tukey [Ref. 8].

The overall objective of LOWESS, like most smoothing or regression routines, is to compute a "fitted" value, $\hat{Y}$, that depicts the middle of the empirical distribution of Y at each X. Unfortunately, most data sets do not contain enough repeated observations at each X to provide a good estimate of the middle of this distribution. LOWESS derives its estimate of $\hat{Y}$ from the equation of a weighted least squares regression line fitted to a set of data points whose X values are located in a user defined neighborhood about Xi (X value of the point being smoothed).

B. MATHEMATICAL DETAILS: NON-ROBUST LOWESS SMOOTHING

The first step in generating a LOWESS smoothed point consists of forming a neighborhood, Figure 2.1, centered around Xi and comprised of its Q nearest neighbors. The user

[1] A brief theoretical explanation of these techniques was presented in Chapter I.

determines Q by choosing the parameter F, which is approximately equal to the percentage of the number of data points used in computing each fitted value. Q is (F x N) rounded to the nearest integer, and the Q nearest neighbors are those points whose X values are closest to Xi. Note that there are not necessarily an equal number of neighborhood points on either side of Xi. Also, Xi is considered to be a neighbor of itself. The parameters F and Q, determined prior to smoothing the first data point, are held constant and used throughout the procedure.

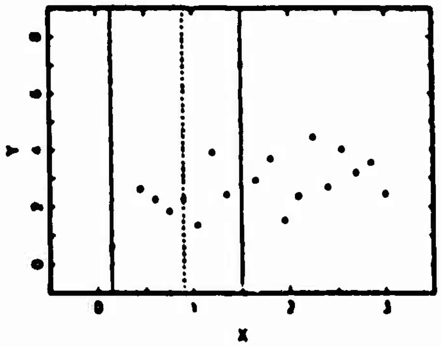


Figure 2.1 Vertical Strip Containing the 10 Nearest Neighbors of X6 in Data Set Two.

In Figure 2.1, the point to be smoothed, X6, is highlighted by a dotted line and the strip boundaries are delineated by solid lines passing through X1 and X10.

STEP TWO consists of defining the local weighting function and calculating individual weights for each point, (Xk,Yk), in the strip formed during STEP ONE. This weighting function is to be centered at Xi and scaled so that it hits zero for the first time at the Q^{th} nearest neighbor of Xi (the strip boundary furthest from Xi). Functions having the following properties will satisfy these requirements:

1. $W(U) > 0$ for $|U| < 1$ (positivity),
2. $W(-U) = W(U)$ (symmetry),
3. $W(U)$ is a nonincreasing function for $u > 0$,
4. $W(U) = 0$ for $|U| > 1$.

Cleveland, [Ref. 3], suggests using a tricube weight function of the form:

$$W(U) = \begin{cases} (1 - |U|^3)^3 & \text{FOR } |U| < 1 \\ 0 & \text{OTHERWISE} \end{cases}$$

Note that this function uses the absolute value of U. The weight given to any point within the strip is calculated by:

$$W(U) = W\left[\frac{X_I - X_K}{D_I}\right]$$

The variable Di is the distance along the X axis from Xi to its $Q^{\underline{th}}$ nearest neighbor. This is the distance from X6 to the left hand boundary in Figure 2.1. When LOWESS starts its smoothing pass at X1, the right hand boundary passes through its $Q^{\underline{th}}$ nearest neighbor, X10 in this example. The neighborhood which, at that time, contains the points X1 ... Xq remains fixed until the distance (Xi-X1) is greater than (Xq-Xi). This usually occurs at i = Q/2 for evenly spaced data. At this point the neighborhood is advanced and the Q nearest neighbor shifts to the left hand boundary where it remains until all of the data points have been smoothed. Di therefore, is generally the distance from Xi to the right hand boundary for i = 1...(Q/2) and is the distance from Xi to the left hand boundary for i = (Q/2)...N.

The weight given to any point in the strip is equal to the height of the curve, W(u), at Xk, Figure 2.2. This figure demonstrates that the tricube weight function:

1. gives the largest weight to the point being smoothed;
2. decreases smoothly as Xk moves away from Xi;
3. is symmetric about the point being smoothed;
4. hits zero for the first time at the Q^{th} nearest neighbor of Xi.

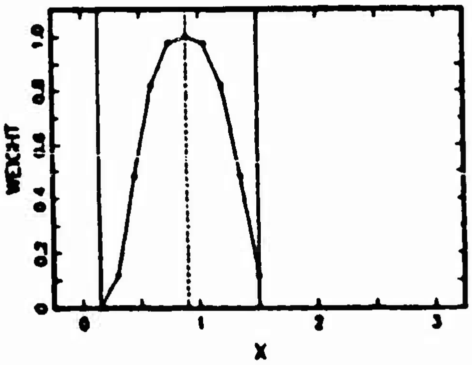


Figure 2.2 TRICUBE Weight Function for the 10 Nearest Neighbors of X6 in Data Set Two.

In cases where several points have abscissas equal to Xi, all of them are given weight 1. If Di is zero, meaning that all Q points in the strip have abscissas equal to Xi, it is impossible to estimate the slope of a fitted line. In this instance, a constant equal to the mean Y value for all Q points is fitted to the point (Xi,Yi).

STEP THREE uses weighted least squares regression to fit a polynomial of degree P to the data points that lie within the strip containing Xi. The parameters of the equation that describes this line are the values of Bj j = 0,1,...P that minimize:

$$\sum_{K=1}^{Q} W_K(U)(Y_K - B_0 - B_1 X_K - \dots B_P X_K^{P})^2$$

Figure 2.3 shows straight (p=1) and quadratic (p=2) lines fit to the neighborhood points surrounding X6 in data set two.

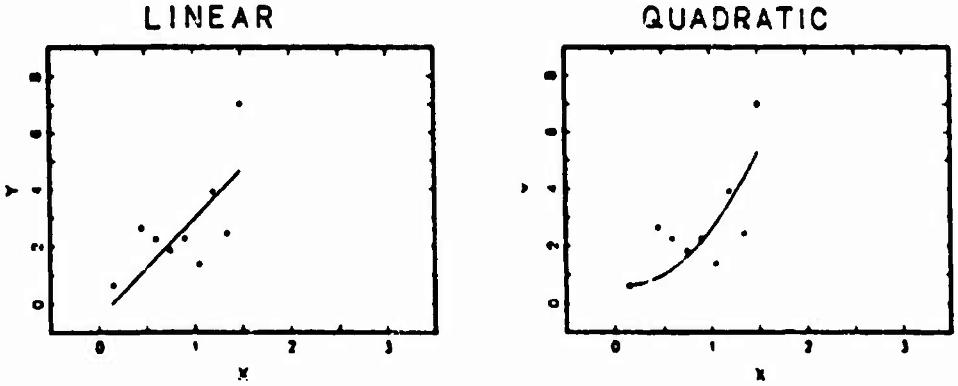


Figure 2.3 Linear and Quadratic Fits.

The choice of an appropriate P depends on the user's perception of the relationship between the points within each neighborhood, the need for flexibility to reproduce patterns in the data, and computational ease. The existence of physical theories that define the relationships as being nonlinear might also influence this choice. Smoothed curves based on higher order polynomial regressions tend to follow abrupt pattern changes better than those based on linear models. Cleveland [Ref. 3], feels that computational considerations begin to override the need for flexibility for values of P greater than 1.

The smoothing routine written for this thesis is capable of performing linear or quadratic regressions. Using p = 1 or 2 should provide adequately smoothed points for any data set.

The final step in the Locally Weighted Regression portion of the LCWESS procedure is the determination of the smoothed point $(Xi,\hat{Y}i)$, Figure 2.4, where:

$$\hat{Y}_i = \sum_{J=1}^{p} B_J(X_i)\cdot X_i^J$$

The notation used here emphasizes that the coefficients of the X_i^J are different for each point Xi.

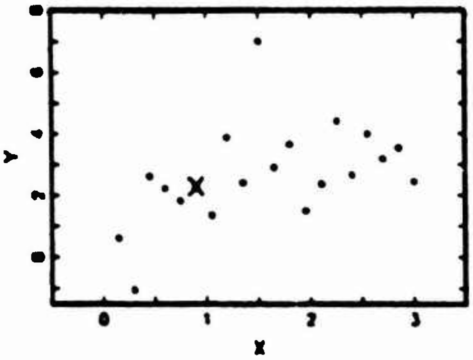


Figure 2.4 Scatter Plot of Data Set Two Superimposed With Smoothed Point (X6,Y6).

LOWESS differs from most other smoothing routines because it smooths all of the data points. This becomes important when smoothing small data sets, when important pattern changes take place near the ends of the data set, or when the smoothed curve is to be used as a regression line to predict future trends. Figure 2.5 summarizes the sequence of steps described above, as they are used to compute a "fitted" value for (X20,Y20), the right hand end point in data set two.

A comparison of Figures 2.1 and 2.5 reveals that the widths of the vertical strips about (X6,Y6) and (X20,Y20) are not equal. Note that the ten nearest neighbors of X20 are all to the left. Although both strips contain ten data points, the requirement to center them around their

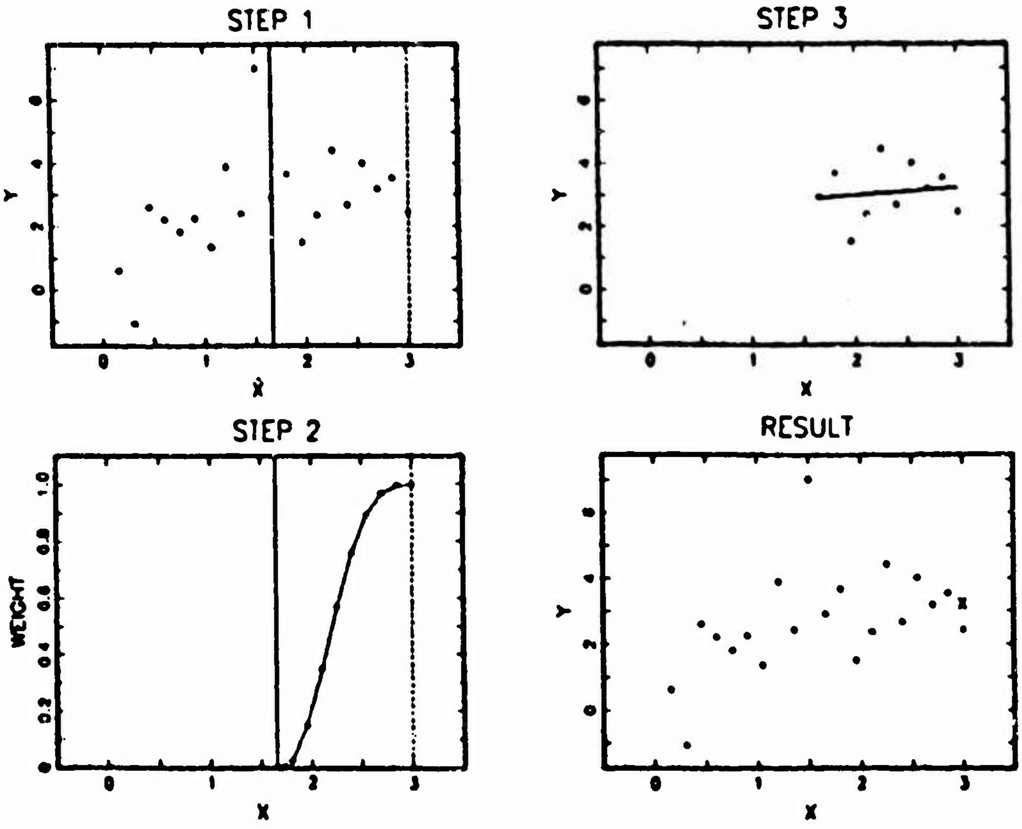


Figure 2.5 Summary of Steps Required for Computing the Smoothed Value at (X20,Y20) in Data Set Two.

respective (Xi,Yi) points forces the right hand portion of the weighting function in Figure 2.5 to fall off-scale. The left hand portion of the weighting function for (X1,Y1) is forced off scale for the same reason. These partial weighting functions still fulfill all of the requirements outlined earlier, however. Unequal spacing of the X's also creates variable strip widths.

A set of smoothed data points, Figure 2.6, is obtained by completing the aforementioned steps for each point in the original data set.

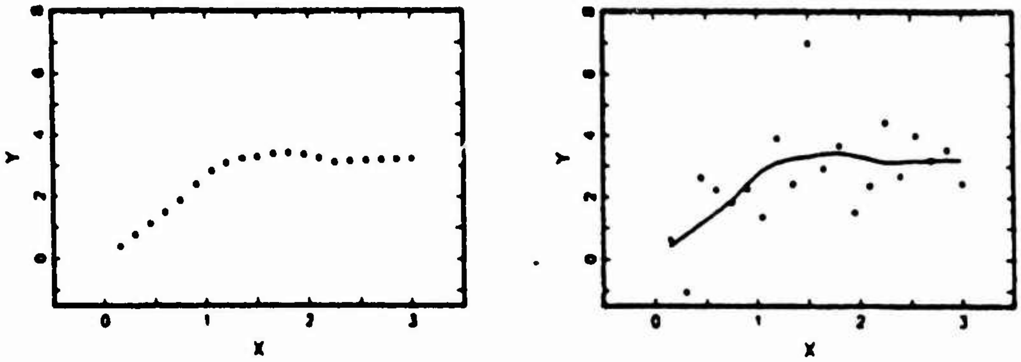


Figure 2.6 Plots of Lowess Smoothed Data Points and Smoothed Curve Superimposed on Data Set Two, (F=.5).

C. MATHEMATICAL DETAILS: ROBUST LOWESS SMOOTHING

The robust smoothing feature of LOWESS prevents a small number of outliers from distorting the smoothed curve. The point (X10,Y10) in Figure 2.1 is one such outlier.

The robust procedure computes a new set of weights for each (Xi,Yi) based on the size of the residuals, $(Yi-\hat{Y}i)$, obtained after the first smoothing pass, Figure 2.7.

Cleveland [Ref. 3], suggests using a bisquare function of the form:

$$D(V) = \begin{cases} (1 - V^2)^2 & \text{FOR } |V| < 1 \\ 0 & \text{OTHERWISE} \end{cases}$$

Robustness weights for each point are calculated by:

$$D_K(V) = D\left[\frac{R_K}{6M}\right]$$

where M is the median of the absolute value of the residuals, Figure 2.8. This is sometimes referred to as the Median Absolute Deviation (MAD).

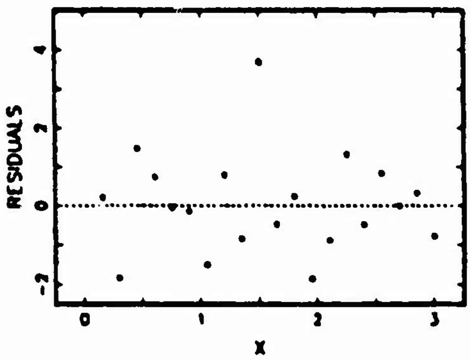


Figure 2.7 Residuals (Yi-Yi) Versus Xi for the Non-Robust Smoothed Points of Data Set Two.

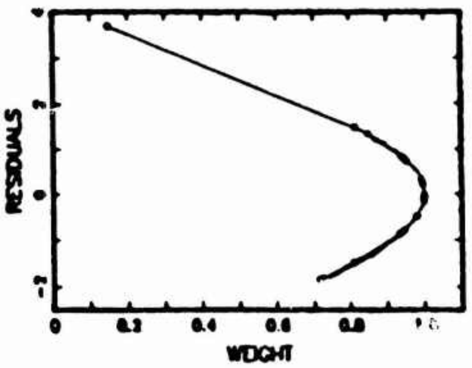


Figure 2.8 Robust Weighting Function For the First Pass Through Data Set Two.

This scheme gives small weights to points associated with large residuals and large weights to points with small residuals. One iteration of the robust locally weighted regression procedure is completed by calculating a new set of "fitted" values using the weighting function

$$WT = W(U) \times D(V)$$

in step three.

Execution of the entire LOWESS algorithm consisting of one locally weighted regression pass and two robust locally weighted regression passes produces a robust smoothed curve, Figure 2.9. The effect of the "outlier" can be seen very clearly.

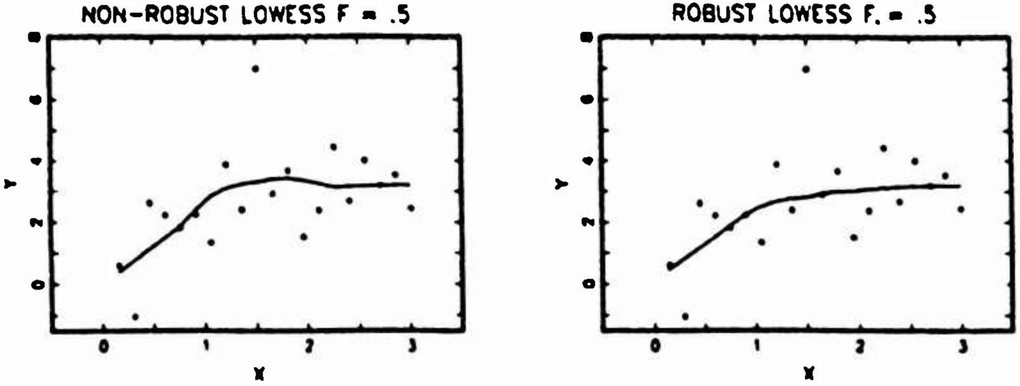
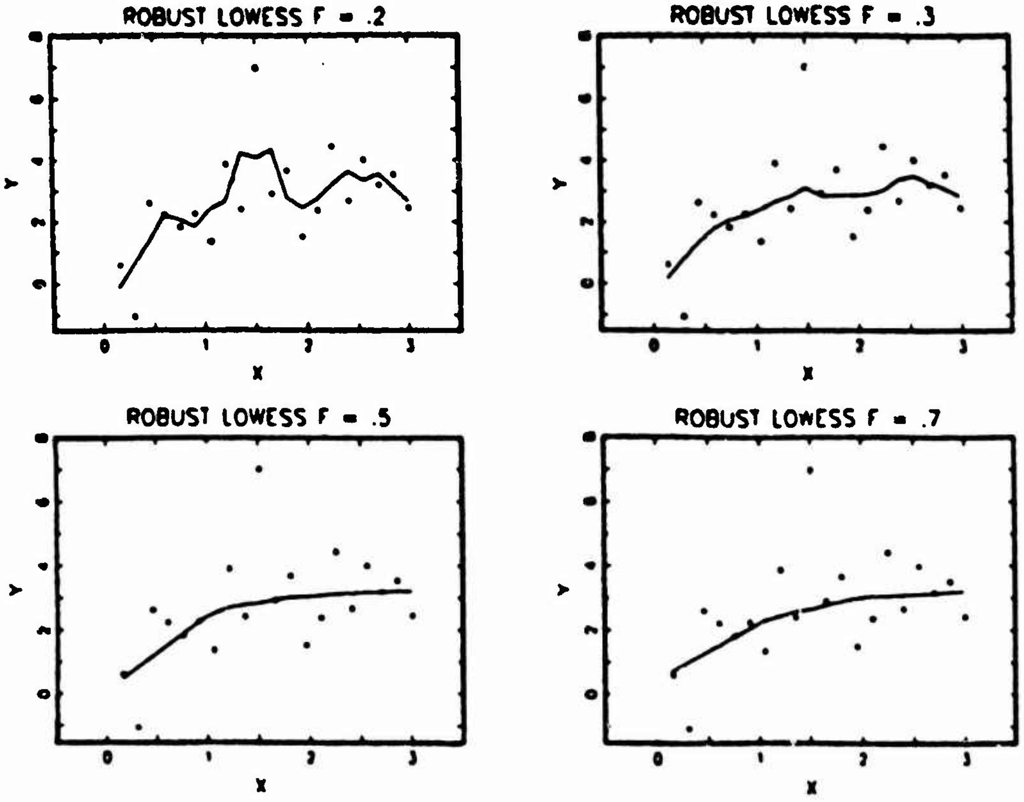


Figure 2.9 Comparison of Non-Robust and Robust LOWESS Smoothing of Data Set Two, (F=.5).

Cleveland [Ref. 3], reports that the number of computations required to complete the LOWESS algorithm on an entire data set is on the order of FN^2. For example, 60 linear regressions were used to complete the robust smoothing of the 20 artificial data points in Figure 2.9. The non-robust curve, on the other hand, required 2/3 fewer calculations and took less than 1/2 the time. The number of calculations required to produce a smoothed curve presents no significant problem for plots of fewer than 100 points. Computational time can be saved by grouping the X_i's on data sets that have repeated X values. This saving results from the fact that if $X_{i+1} = X_i$ then $\hat{Y}_{i+1} = \hat{Y}_i$. Assigning the same Y_i value to each of the N_i repeated X_i's reduces the number of regressions required by N_i for non-robust smoothing and by $3N_i$ for robust smoothing.

D. CHOOSING F

There are no set criteria for choosing F. Small values produce curves with high resolution and a lot of noise. Larger F's produce curves with low resolution and less noise, but require increased computational time. In general, increasing F tends to produce smoother curves, Figure 2.10. Cleveland, [Ref. 3], suggests that values between .2 and .8 should be satisfactory for most purposes. The goal is to choose the largest F that minimizes the variability in the smoothed points without distorting patterns in the data. Computational time may become a consideration in choosing F when smoothing large data sets. In general though, F will decrease as the series length increases.


ROBUST LOWESS F = .2

ROBUST LOWESS F = .3

ROBUST LOWESS F = .5

ROBUST LOWESS F = .7

Figure 2.10 Comparison of Robust LOWESS Smoothing of Data Set Two for Different Values of F.

Smoothing routines, LOWESS included, do not provide regression equations or other analytical results on which to test goodness of fit. The user must judge the adequacy of the results. The choice of F is not so critical for cases in which the purpose of the smoothing is to enhance the visual perception of gross patterns in the data. For example, the rough curve obtained by using F=.2 on data set two, the left hand plot of Figure 2.10, provides an adequate picture of an overall increasing trend. More care must be taken in some applications, such as time series analysis, or when the smoothed (Xi,Yi) values may be used as a type of regression function, or finally, when the smoothed curve may be presented without an accompanying plot of the original data points. Taking F=.5 is a reasonable choice when there is no clear idea of what is needed, [Ref. 3]. Chambers, [Ref. 1], suggests that it is often wise to try several values of F before selecting the "best" one for a particular application.

Techniques for determining bandwidth using techniques of cross-validation have been considered by Cleveland [Ref. 3], and Rice [Ref. 9], but are not included here.

III. EVALUATION OF THE LOWESS CURVE SMOOTHING PROGRAM

A. GENERAL

Smoothing routines are generally used to filter noisy data and approximate underlying relationships that may be too complex to describe mathematically or too difficult to fit by simple polynomial regression. Effective routines must be flexible and local. They must allow the data to determine the shape of the smoothed curve and they must be able to follow abrupt as well as smooth changes in curvature. This evaluation will test LOWESS in each of these areas.

B. METHODOLOGY

LOWESS, like most other curve smoothing schemes, provides no analytical solutions by which to measure its effectiveness. The correctness or adequacy of the fit must be judged subjectively. And there are no standard guidlines to follow. Sometimes the shape of the fit can be checked by comparing it to the physical laws that govern the application at hand. The programs written to support this thesis were evaluated by:

1. examining their performance on a set of test data for which the underlying functional relationships were known;
2. comparing their results with those obtained from widely used and previously validated curve smoothing techniques, namely; LEAST SQUARES REGRESSION, MOVING AVERAGE and COSINE ARCH weighted smoothing.

The theory of moving average procedures dates back to definitive studies of discrete time series models completed by H. Wold in the mid 1930's. The general process is based on the assumptions and theories recounted in Chapter I. The moving average is defined by the expression

$$X(T) = \sum_{J=-M}^{N} A_J \, Z(T-J) \qquad T = 0,\ 1 \ldots$$

where M and N are nonnegative integers and the weighting coefficients A_j are real constants. Kendall and Stuart [Ref. 4], and Koopmans [Ref. 10], present in depth discussions and theoretical derivations that expand on the ideas presented in Chapter I. The moving average routine employed in this analysis is contained in the IBM GRAFSTAT statistical graphics package. The weighting function used in that program takes the form

$$A_J = \frac{1}{M} \qquad J = -M \ldots N$$

The COSINE ARCH smoothing procedure used here, is a moving average process that uses a cosine weighting function of the form

$$A_J = \frac{1}{M+1}\ 1 - \cos\frac{2\pi(J+1)}{M+1} \qquad J = 0, 1 \ldots N-1$$

It is characterized as a good smoother by Anscombe, [Ref. 11], and is often used as a trend remover during time series analysis.

C. TESTING PROCEDURES AND RESULTS

Three sets of test data were developed to check all aspects of the LOWESS program's capabilities; its ability to follow linear trends as well as abrupt and smooth changes in curvature.

1. Phase One: Linear Trends

Test set one, Figure 3.1, consists of 150 data points having the following functional relationship:

$$Y = X + \text{NORMAL}(0,1)\ \text{NOISE} \qquad 0 \le X \le 10$$

was designed to test IOWESS' ability to detect linear trends in noisy data. Although this test appears redundant, many complex smoothing procedures have failed because they did not return straight lines when that was the shape of the underlying curve.

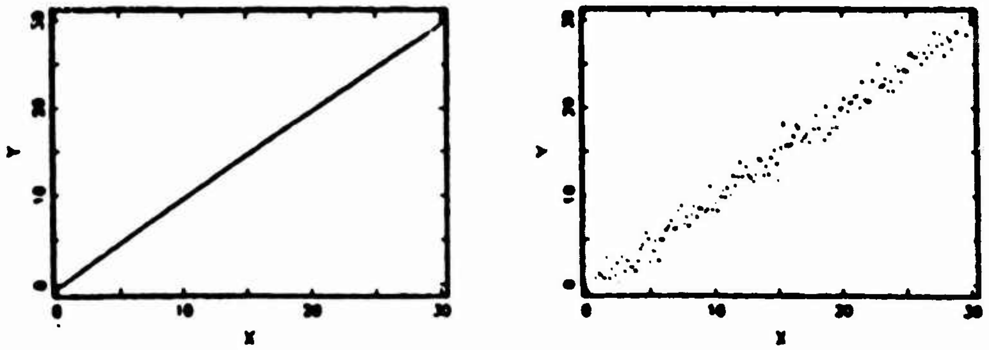


Figure 3.1 Test Set One With and Without N(0,1) Noise.

The adequacy of LOWESS' performance on test set one was measured by comparing it with a linear least squares regression line fitted to the same data.

As pointed out in CHAPTER II, LOWESS produces increasingly smoother curves as the parameter F approaches 1. When F=1, each neighborhood used throughout the smoothing process contains N • 1 = N points. This implies that each smoothed point $(Xi,\hat{Y}i)$ is computed from the equation of the TRICUBE weighted regression line fitted to all of the data. This procedure should produce a LOWESS smoothed curve that closely resembles the linear regression of Y on X. The TRICUBE weighting function used in LOWESS may cause minor disparities between the two "fits," however. A visual inspection of the bottom two plots in Figure 3.2 reveals that LOWESS and the linear regression produced nearly identical "fits."

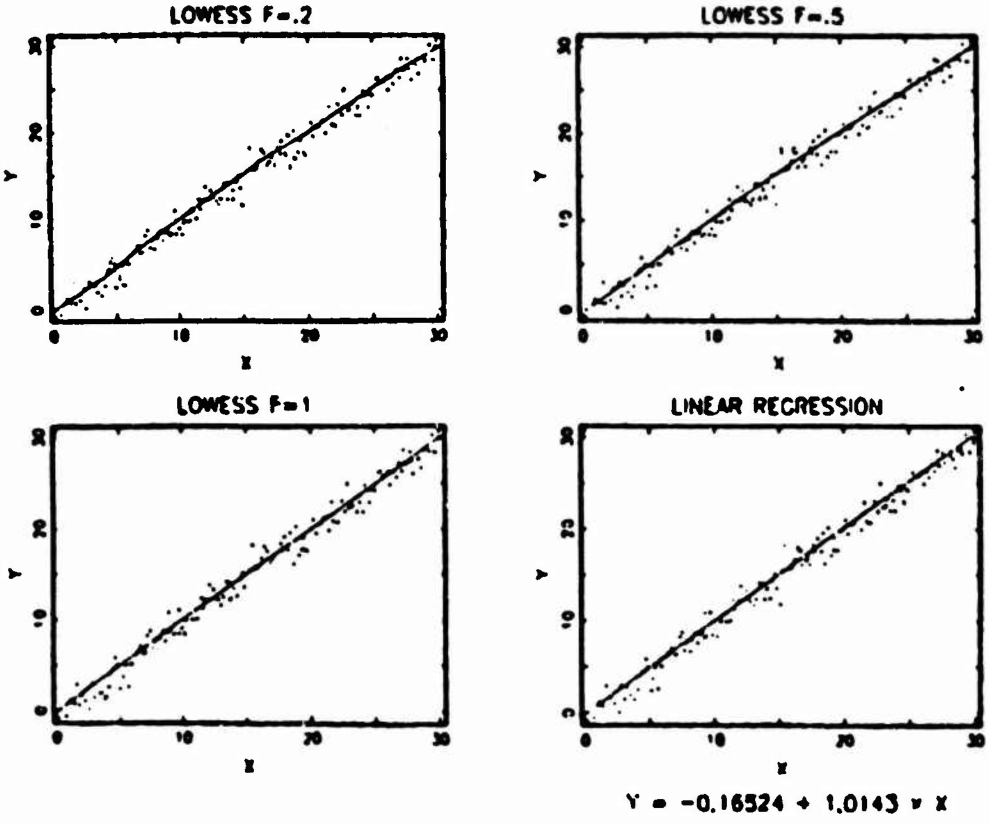


Figure 3.2 Comparison of LOWESS Smoothing and Linear Regression of Test Set One.

Goodness of fit can te measured by examining the residuals $(Yi-\hat{Y}i)$ from each smoothing procedure. A perfect reprcduction of the underlying functional relationship, Y =

X, would produce a set of residuals distributed Normal(0,1), the same distribution found in the noise. The results of the GRAFSTAT distribution fitting proceedure summarized in Table II indicate that the distribution of the regression residuals can be approximated as Normal(0,1.04) while the LOWESS residuals are approximately Normal(.002,1.016).

Hypothesis tests comparing the means and variances of these distributicns with those of the Normal(0,1) distributed noise, will provide some measure of the goodness of fit of each smoothing scheme. The results of these tests, conducted at the 95% confidence level, are summarized in Table I.

The output of the GRAFSTAT distribution fitting procedure presented in Table II and the hypothesis tests summarized in Table I, suggest that there is no significant difference between the distribution of the residuals from the linear regression or LOWESS smoothing of test set one, and the Normal(0,1) noise inccrporated into the data. This provides strong support for the premise that LOWESS depicts linear trends very well. Visual comparison of the LOWESS smooths in Figure 3.2 confirms that LOWESS follows the same general trend regardless of what F is used; small values provide rougher curves that have the same general slope.

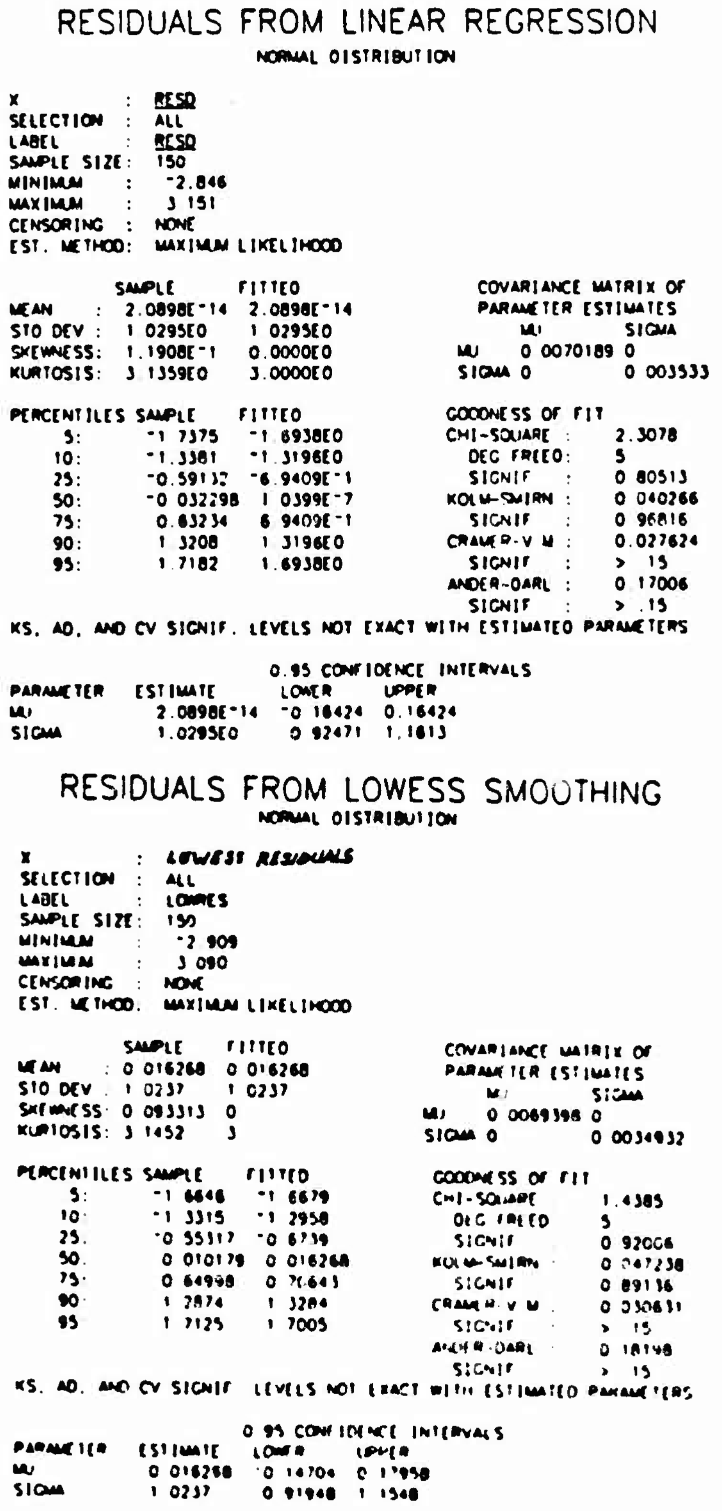
TABLE I

Comparison of the Means and Variances of Residuals
From Smooths of Test Set One to the Normal(0, 1) Noise

			ncise	T	$Z(1-\alpha/2)$		β
linear	mean	0.000	0	0.000	1.96	accept	0.05
rgrsn	var	1.040	1	0.346	1.96	accept	0.07
LOWESS	mean	0.002	0	0.024	1.96	accept	0.05
	var	1.016	1	0.138	1.96	accept	0.06

TABLE II

Summary of GRAFSTAT Distribution Fitting of Residuals from Regression and LOWESS Smooths of Test Set One



```
                 RESIDUALS FROM LINEAR REGRESSION
                        NORMAL DISTRIBUTION

X           :  RESD
SELECTION   :  ALL
LABEL       :  RESD
SAMPLE SIZE:   150
MINIMUM     :   ¯2.846
MAXIMUM     :   3.151
CENSORING   :  NONE
EST. METHOD:   MAXIMUM LIKELIHOOD

            SAMPLE        FITTED                  COVARIANCE MATRIX OF
MEAN    :   2.0898E¯14    2.0898E¯14              PARAMETER ESTIMATES
STD DEV :   1.0295E0      1.0295E0                      MU           SIGMA
SKEWNESS:   1.1908E¯1     0.0000E0              MU     0.0070189 0
KURTOSIS:   3.1359E0      3.0000E0              SIGMA  0             0.003533

PERCENTILES SAMPLE        FITTED               GOODNESS OF FIT
     5:         ¯1.7375     ¯1.6938E0          CHI-SQUARE  :    2.3078
    10:         ¯1.3381     ¯1.3196E0            DEG FREED:     5
    25:         ¯0.59132    ¯6.9409E¯1           SIGNIF    :    0.80513
    50:         ¯0.032298    1.0399E¯7         KOLM-SMIRN  :    0.040266
    75:          0.63234     6.9409E¯1           SIGNIF    :    0.96816
    90:          1.3208      1.3196E0          CRAMER-V M  :    0.027624
    95:          1.7182      1.6938E0            SIGNIF    :    > .15
                                               ANDER-DARL  :    0.17006
                                                 SIGNIF    :    > .15
KS, AD, AND CV SIGNIF. LEVELS NOT EXACT WITH ESTIMATED PARAMETERS

                          0.95 CONFIDENCE INTERVALS
PARAMETER    ESTIMATE       LOWER      UPPER
MU             2.0898E¯14   ¯0.16424   0.16424
SIGMA          1.0295E0      0.92471   1.1813

                 RESIDUALS FROM LOWESS SMOOTHING
                        NORMAL DISTRIBUTION

X           :  LOWESS RESIDUALS
SELECTION   :  ALL
LABEL       :  LOWRES
SAMPLE SIZE:   150
MINIMUM     :   ¯2.909
MAXIMUM     :   3.090
CENSORING   :  NONE
EST. METHOD.   MAXIMUM LIKELIHOOD

            SAMPLE      FITTED                  COVARIANCE MATRIX OF
MEAN     :  0.016268    0.016268                PARAMETER ESTIMATES
STD DEV  .  1.0237      1.0237                        MU           SIGMA
SKEWNESS:   0.093313    0                     MU     0.0069398 0
KURTOSIS:   3.1452      3                     SIGMA  0             0.0034932

PERCENTILES SAMPLE      FITTED               GOODNESS OF FIT
     5:         ¯1.6646    ¯1.6679           CHI-SQUARE        1.4385
    10:         ¯1.3315    ¯1.2958             DEG FREED       5
    25.         ¯0.55317   ¯0.6739             SIGNIF          0.92006
    50.          0.010179   0.016268         KOLM-SMIRN .      0.047238
    75:          0.64998    0.70641            SIGNIF          0.89136
    90:          1.2874     1.3284           CRAMER-V M .      0.030631
    95           1.7125     1.7005             SIGNIF   .      > .15
                                             ANDER-DARL .      0.18198
                                               SIGNIF          > .15
KS. AD. AND CV SIGNIF  LEVELS NOT EXACT WITH ESTIMATED PARAMETERS

                        0.95 CONFIDENCE INTERVALS
PARAMETER    ESTIMATE     LOWER      UPPER
MU           0.016268    ¯0.14704    0.17958
SIGMA        1.0237       0.91948    1.1548
```

2. Phase Two: Abrupt Changes in Curvature

Test set two, Figure 3.3, consisting of 220 data points having the following mathematical relationship

$$Y = \begin{cases} .4X + \text{NORMAL}(0,1)\ \text{NOISE} & 0 \le X \le 10 \\ 3 + .1X + \text{NORMAL}(0,1)\ \text{NOISE} & 10 < X \le 25 \\ 14.6 - 3.67X + \text{NORMAL}(0,1)\ \text{NOISE} & 25 < X \le 40 \\ 0 + \text{NORMAL}(0,1)\ \text{NOISE} & 40 < X \le 44 \end{cases}$$

was used to test LOWESS' ability to handle abrupt pattern changes. The smooth of test set two generated by LOWESS, was compared to those produced by MOVING AVERAGE and COSINE ARCH filtering of the same data.

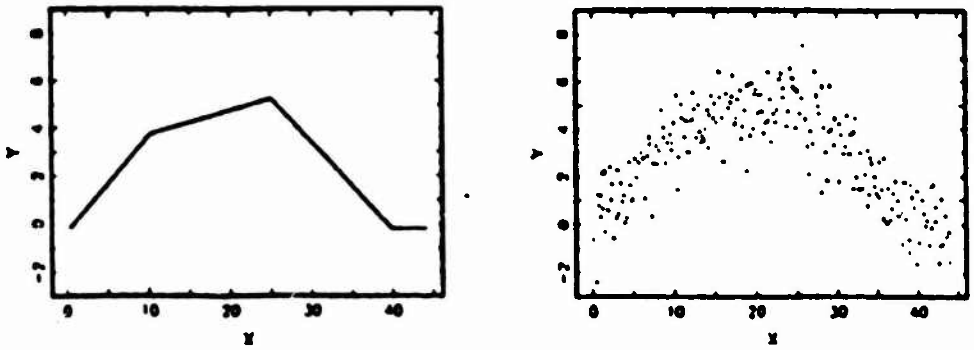


Figure 3.3 Test Set Two With and Without N(0,1) Noise.

Determining the amount of smoothing required by a data set is, perhaps, the most difficult aspect of using any curve smoothing routine. Smoothness is controlled by the size of the parameter F in LOWESS and by the parameter M (bandwidth) in MOVING AVERAGE and COSINE ARCH smoothing. These parameters determine the number of points, or neighborhood size, used to compute each smoothed value. The goal, regardless of the method chosen, is to use the largest neighborhood that minimizes the variability in the smoothed points without distorting patterns in the data. Another factor that must also be considered when choosing M, is that MOVING AVERAGE and COSINE ARCH smoothing routines produce only (N-M) smoothed points. Using proportionately large values of M, therefore, might result in losing significant portions of the original pattern at the ends. This shortcomming will be evident in the graphical comparisons made throughout the remainder of this chapter.

Comparison tests made during phases two and three of this evaluation used selected LOWESS smooths and corresponding MOVING AVERAGE and COSINE ARCH smoothed curves. Parameters for the three processes are directly convertible by the relationship $M = F \cdot N$.

Figure 3.4 presents graphical comparisons of LOWESS smooths (solid line) using parameter values F = .15,.25,.50 and .75 to illustrate some of the considerations made during the parameter selection phase of
a smoothing operation. The exact underlying relationships (dashed lines) were included to demonstrate how large values of F can cause pattern distortion.

It is apparent from the sequence of illustrations in Figure 3.4, that LOWESS produces smoother curves as F increases. The smoothest curves are not always the most desireable, however. The bottom two curves (F=.50 and F=.75) have distorted the original pattern by using too many points to compute the smoothed values. Test set two contains 50 points in the segment $(0 \leq X \leq 10)$. Using a neighborhood much larger than $220 \cdot .25 = 55$ points on this data set would have a tendency to fit the wrong slope to the first linear segment. Additionally, it would cause over smoothing of the corners. Figure 3.5 shows the neighborhood and linear regression used to smooth the point (X10,Y10) during production of the smoothed curve (F=.75) pictured in the lower right corner of Figure 3.4. It is easy to see that following this slope would distort the pattern presented by the data.

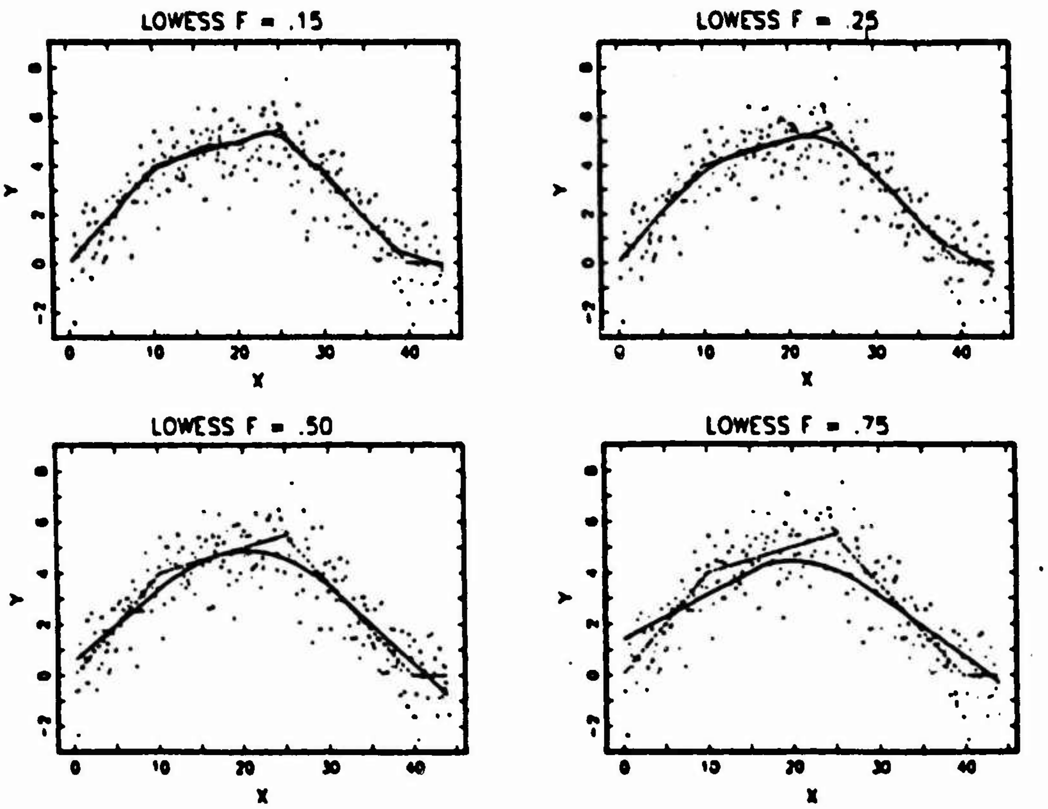


Figure 3.4 Comparison of LOWESS Smoothing of Test Set Two Using Different Values of the Parameter F.

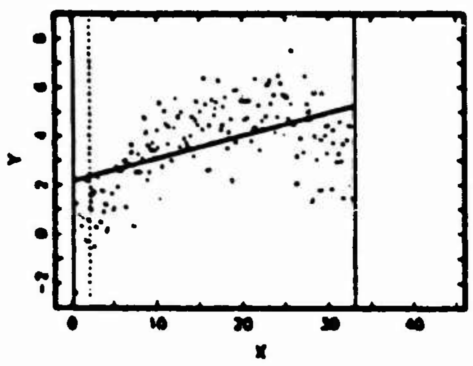


Figure 3.5 Linear Regression Step in Smoothing (X10,Y10) in Test Set Two Using LOWESS With F=.75.

The F=.15 plot depicted in Figure 3.4, demonstrates that small F's create very locally smoothed curves that contain a great deal of noise but follow gross patterns very well. Using a small F is an excellent idea if the sole purpose of the smoothing is to highlight major trends in the data.

The LOWESS smoothed curve obtained by using F=.25 is the one best suited for comparison with corresponding MOVING AVERAGE and COSINE ARCH smooths, Figure 3.6.

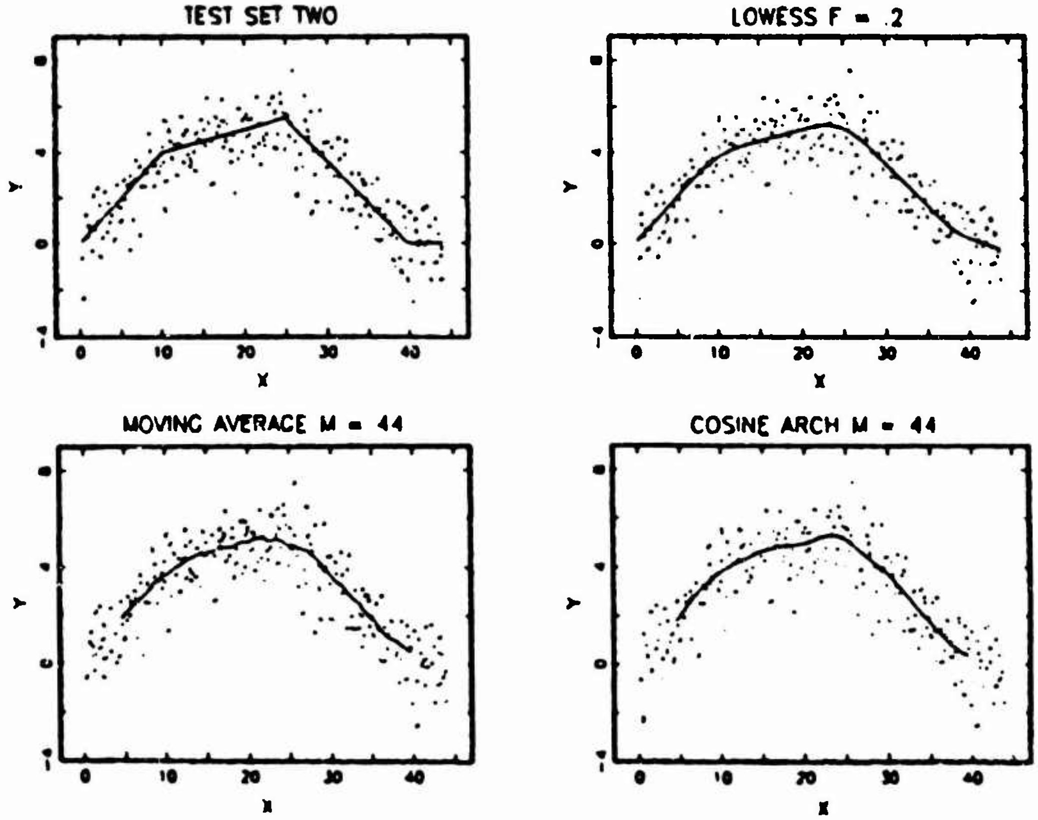


Figure 3.6 Comparison of LOWESS, MOVING AVERAGE and COSINE ARCH Ssmoothing of Ttest Sset Two.

Inspection of the plots in Figure 3.6 reveals that all of the smoothing procedures fit similarly shaped curves to most of the data. The inability of the MOVING AVERAGE and COSINE ARCH routines to smooth the extreme edges of a plot precluded them from fitting a curve to the last segment of test set two. Practitioners of these routines often extend the curve or fit the ends by eye. Applying these techniques to the bottom curves in Figure 3.6 does not reveal any significant pattern changes. LOWESS, although it does not follow the level trend accurately, does reveal a major pattern change in the last section of the data.

All three of the procedures have a tendency to round sharp corners as the parameters F and M are increased. The MOVING AVERAGE curve, in the lower left, has a very rounded shape and does not highlight the linear trend in segments one or two. The COSINE ARCH filter does a little better. It portrays the linearity of section three with nearly the correct slope but fits segments one and two with one smooth curve. Additionally, it has added a misleading hump at the intersection of segments two and three. LOWESS is the only procedure that clearly pictures the underlying pattern as a series of straight lines. An experienced user who understands that LOWESS rounds corners, could almost duplicate the original pattern by connecting the linear portions of the curve.

Smoothing procedures are not only judged on their ability to depict patterns, but are also rated on their ability to filter out unwanted noise. Gross differences in their capabilities can be picked out easily in a graphical comparison. It is readily apparent that the MOVING AVERAGE curve in Figure 3.6 is much noisier that either the LOWESS or COSINE ARCH smooths.

A more analytical measure of a procedure's smoothing ability can be made by comparing periodograms of the unfiltered and filtered data. A periodogram is an analysis technique used to estimate the spectral density function of a time series at periodic frequencies, λ_v. The periodogram function is defined by

$$I_{N,V} = \frac{1}{2\pi N}\left|\sum_{T=1}^{N} X(T)\; E^{-i\lambda_V T}\right|^{2}$$

Refer to Koopmans [Ref. 10], chapter 8, for a detailed discussion of the periodogram and its distributional properties. The periodograms in Figure 3.7 provide

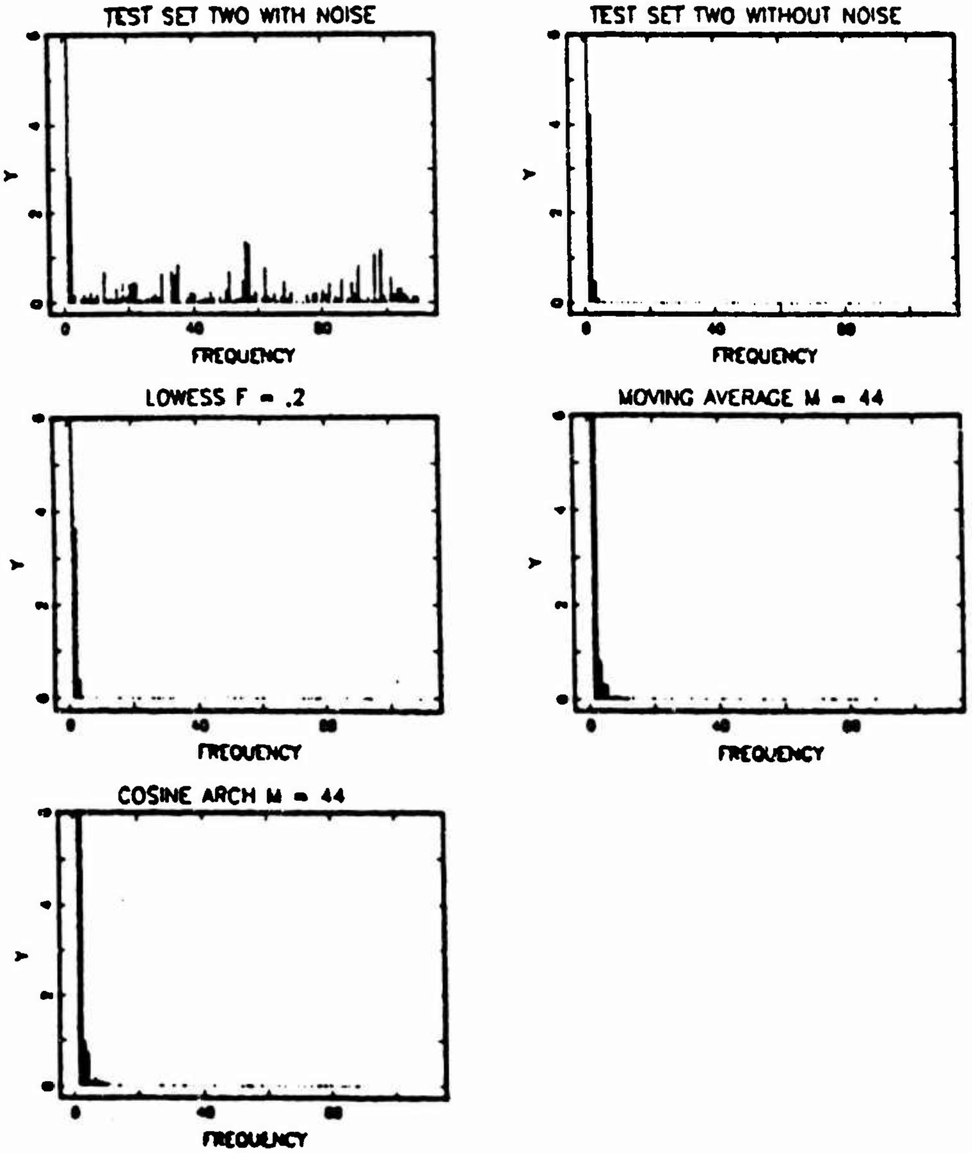


Figure 3.7 Comparison of Periodograms of LOWESS, MOVING AVERAGE and COSINE ARCH Smoothing of Test Set Two.

comparisons of the filtering properties of each smoothing routine. The vertical lines on each plot represent periodicities, the spectral frequencies of which are measured along the abscissa. The height of the lines is an indicator of the significance of the associated frequencies. The plots in Figure 3.7, were truncated at Y = 6 to prevent the obscuration of the minor frequencies.

A visual inspection of these periodograms reveals that LOWESS produces the smoothest (most noise free) curve. In fact, the periodogram of the LOWESS curve and noise free data are nearly identical.

All of this evidence supports the conclusion that LOWESS performs at least as well on data sets that contain abrupt changes in curvature as do the widely accepted MOVING AVERAGE and COSINE ARCH procedures.

3. Phase Three: Smooth Changes in Curvature

Test set three, Figure 3.8, comprised of 100 data points having the following relationship

$$Y = \text{SIN } X + \text{NORMAL}(0,1) \text{ NOISE} \quad 0 \le X \le 2$$

was used to evaluate LOWESS' ability to follow smooth changes in curvature. The same procedures used in the preceding section to test LOWESS' ability to handle abrupt pattern changes were applied here.

Test set three appears to either have a negative linear trend, or appears to cycle about the line Y = 0. A series of LOWESS smooths, Figure 3.9, starting with a small F parameter, was used to discover the general pattern (dashed line) and refine the resulting smoothed curve (solid line). The distorted smooth in the lower right hand plot demonstrates the inherent danger in selecting a large F if only one smoothing pass is planned.

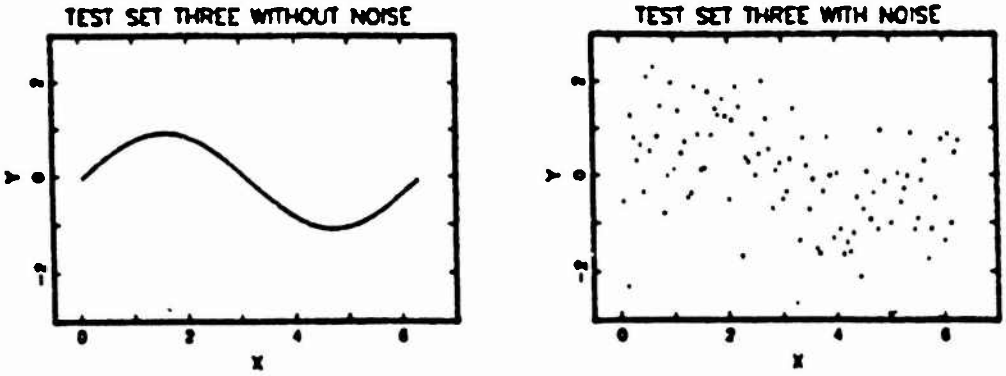


Figure 3.8 Test Set Three With and Without N(0,1) Noise.

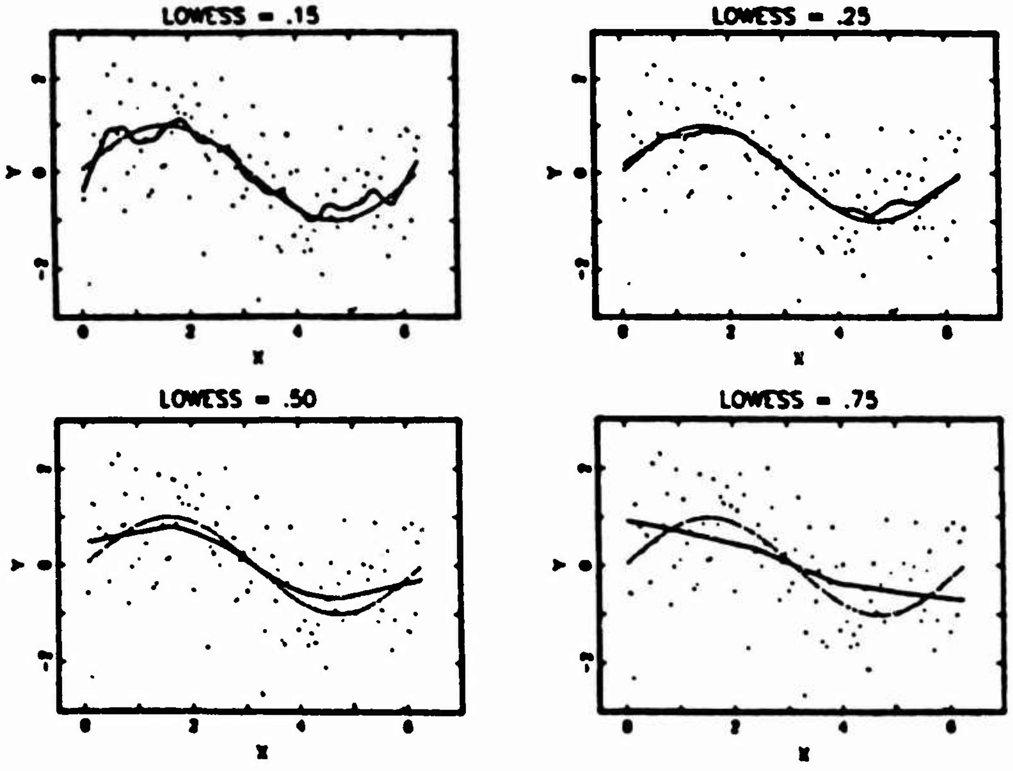


Figure 3.9 Comparison of LOWESS Smoothing of Test Set Three Using Different Values of the Parameter F.

The LOWESS curve obtained by using F=.25 provided the most smoothing without distorting the pattern and was used in a direct comparison with corresponding MOVING AVERAGE and COSINE ARCH smooths, Figure 3.10. The LOWESS smooth is the only curve that has the characteristic sinusoidal shape. The MOVING AVERAGE plot, although very noisy, would present the proper picture if the ends of the curve were extended. The radical change in curvature on the left end of the COSINE ARCH smoothed curve detracts from its abiliity to represent the true shape of test set three.

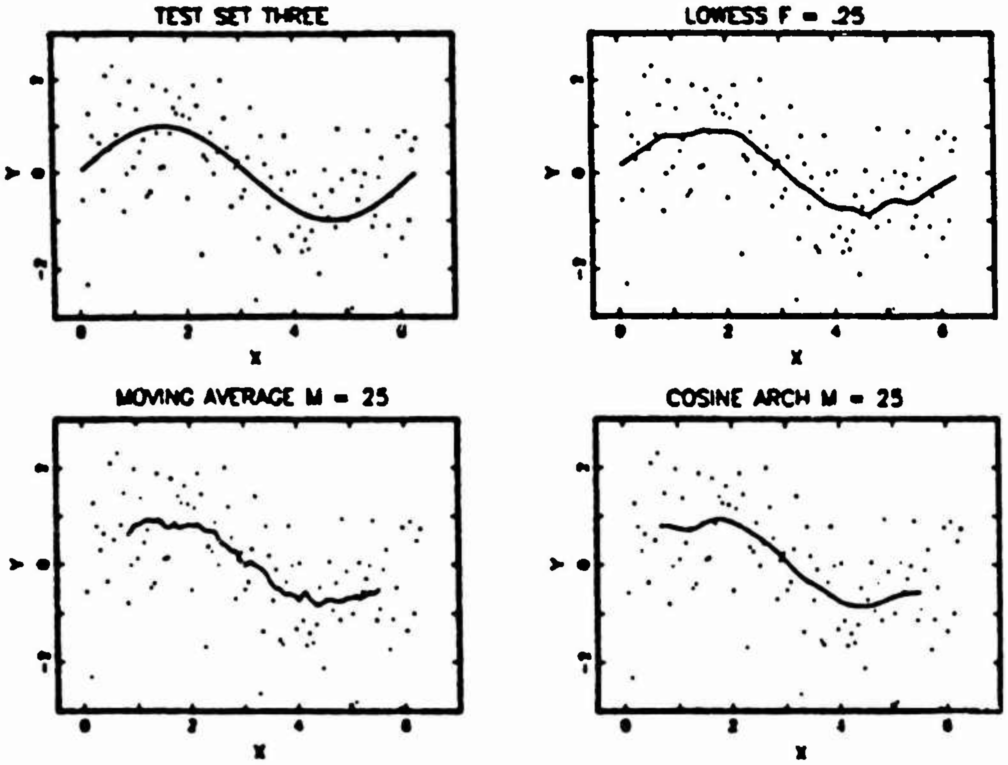


Figure 3.10 Comparison of LOWESS, MOVING AVERAGE and COSINE ARCH Smoothing of Test Set Three.

Comparison of the periodograms presented in Figure 3.11, shows, once again, that LOWESS produces the smoothest curve, while Figure 3.10 shows that it seems to follow the model the best.

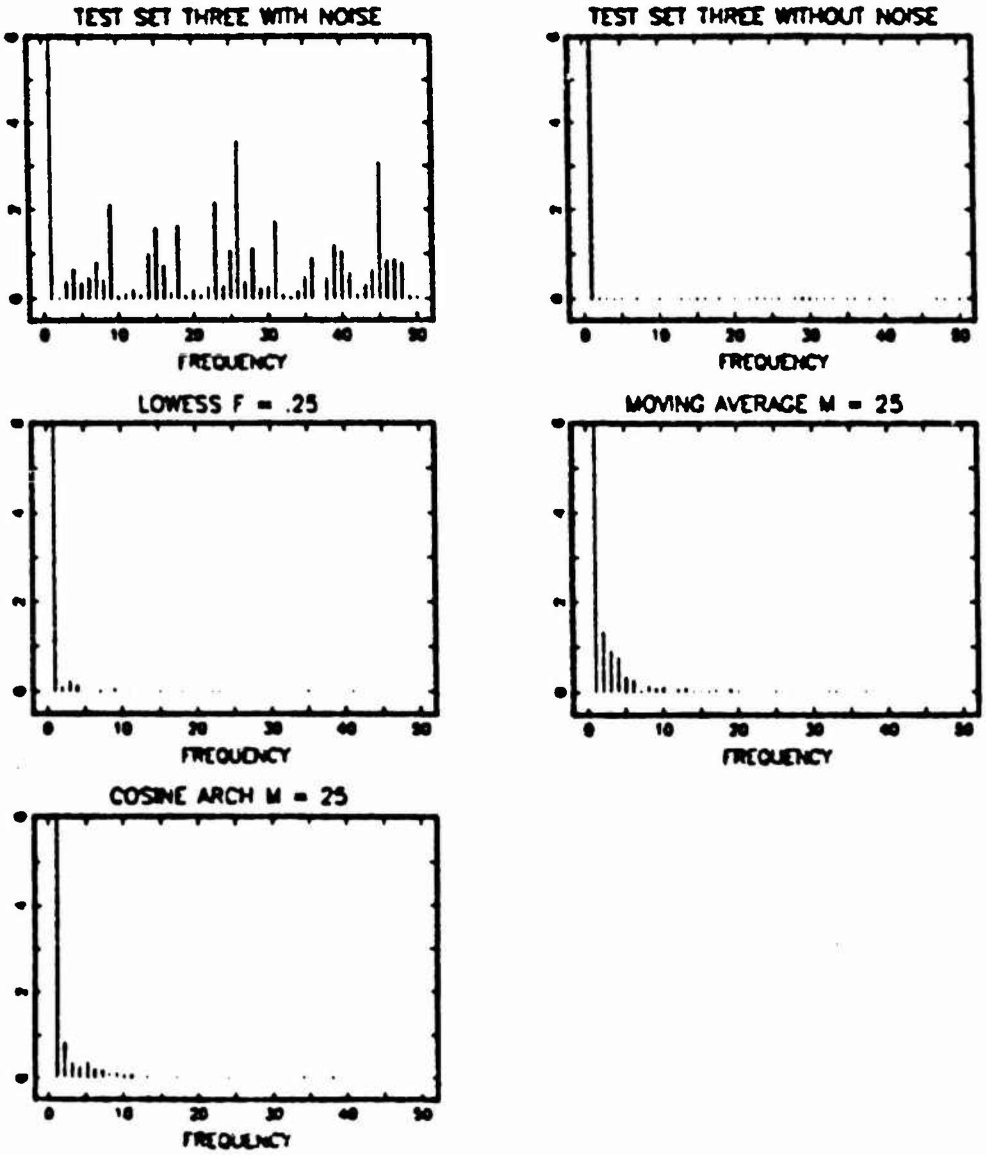


Figure 3.11 Comparison of Periodograms of LOWESS, MOVING AVERAGE ans COSINE ARCH Smoothing of Test Set Three.

The graphical comparisons made in Figure 3.10 and 3.11 demcnstrate clearly that LOWESS performs at least as well as MOVING AVERAGE and COSINE ARCH routines when smoothing data that has a smooth curvilinear pattern.

4. Phase Four: Unequal Spacing

Besides being able to smooth all of the data points, LOWESS enjoys another possible advantage over MOVING AVERAGE type procedures, in that it was designed to work on unequal as well as equally spaced data. The definition of MOVING AVERAGES

$$Y_I = \sum_{J=-M}^{M} A_J Y_{I-J} \qquad I = 0,1,2,\ldots$$

holds only if the Yi's are equally spaced and have a linear relationship over the interval (i-m) ... (i+m). Violation of the linearity assumption introduces bias into the results while violation of the equal spacing requirement invalidates them. LOWESS would indeed enjoy a distinct advantage over MOVING AVERAGE type smoothing procedures if it produces acceptable results on irregularly spaced data.

This section examines LOWESS' ability to smooth two different sets of this of type data. The first, natural log of energy dissipation versus depth, Figure 3.12, is a transformed portion of data collected during a turbulence measuring experiment conducted by the Department of Oceanography, U.S. Naval Postgraduate School.

The LOWESS curves obtained by using linear and quadratic regressions during Step Three of the smoothing procedure were compared to a quadratic least squares regression line fit to the same data, Figure 3.13
Higher order regressions were rejected as plausible solutions because the regression coefficients Bj, j = 3,4,5... were found to be statistically insignificant compared to the Bj, j = 0,1,2 constants. A quadratic relationship also seemed to be a reasonable assumption since turbulence is a

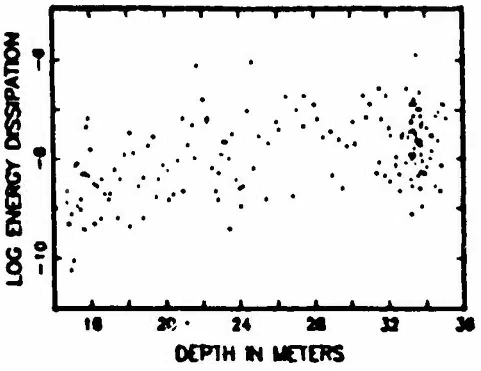
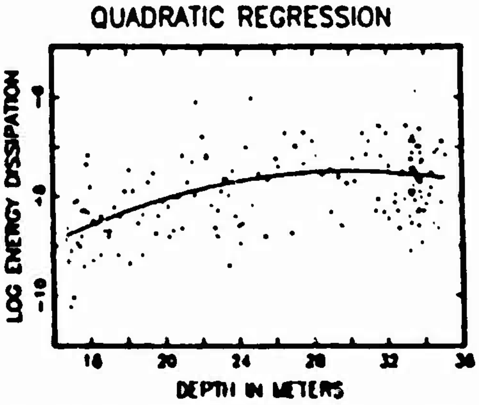


Figure 3.12 Natural Log of Energy Dissipation vs Depth.



```
Y = +/C × X • 0 1 2   WHERE: C = -12.512 0.33412 -0.0055612

                       ANALYSIS OF VARIANCE TABLE

  SOURCE                      SS            DF          MS           F

 GRAND MEAN (SEE NOTE)     10275.656         1
 REGRESSION                   28 970         2        14.485      32.500
 RESIDUAL                     73.094       184          .446
 TOTAL                     10377.719       187        82.142

 THE SIGNIFICANCE LEVEL OF REGRESSION  =  .0000
 (SIGNIFICANCE LEVEL = AREA UNDER CURVE BEYOND COMPUTED F)
 R SQUARE (SEE NOTE)                   =    .284

  NOTE: IN WEIGHTED CASE, SEE DESCRIPTION FOR MEANING
```

Figure 3.13 Quadratic Regression and Analysis of Variance Table for Ln Energy Dissipation Versus Depth.

function of pressure which varies in proportion to depth squared.

Figure 3.14 shows that the LOWESS curves (solid lines) for the linear (P = 1) smooths follow the general quadratic regression (dashed lines) for small values of F but flatten the pattern for large F's. The quadratic (P = 2) LOWESS curves close in on the regression line as F increases and produce a fairly good match as F reaches .75.

The quadratic LOWESS curve also appears to follow local peaks and valleys more accurately for small F's than does its linear counterpart. This is not unexpected. Figure 3.15 shows that the characteristically bowed shape of a quadratic curve produces larger $\hat{Y}i$ values in the middle of a data set (Xi is located in the middle of the LOWESS neighborhood) than a straight line fitted to the same data.

The "fits" of Figure 3.14 can be compared analytically, as was done in the Phase One test, by examining the distribution of their residuals. Combining these analytical results with graphical comparisons provides some goodness of fit measure for the two curves. The nonparametric Smirnov two sample test [Ref. 12], is appropriate in this case because the distribution of the residuals is unknown. The results of this test conducted at the 95% confidence level, Table III, indicate the there is no significant statistical difference between the F=.75 quadratic LOWESS curve and the quadratic least squares regression line. See the lower right hand plot of Figure 3.14

This example demonstrates that LOWESS works quite well on unequally spaced data. It also shows that quadratic LOWESS works better than the linear model when neighborhood sizes are too large to support the assumption that the neighborhood points are related linearly. Quadratic LOWESS should be used whenever the data suggests that that assumption is not true.

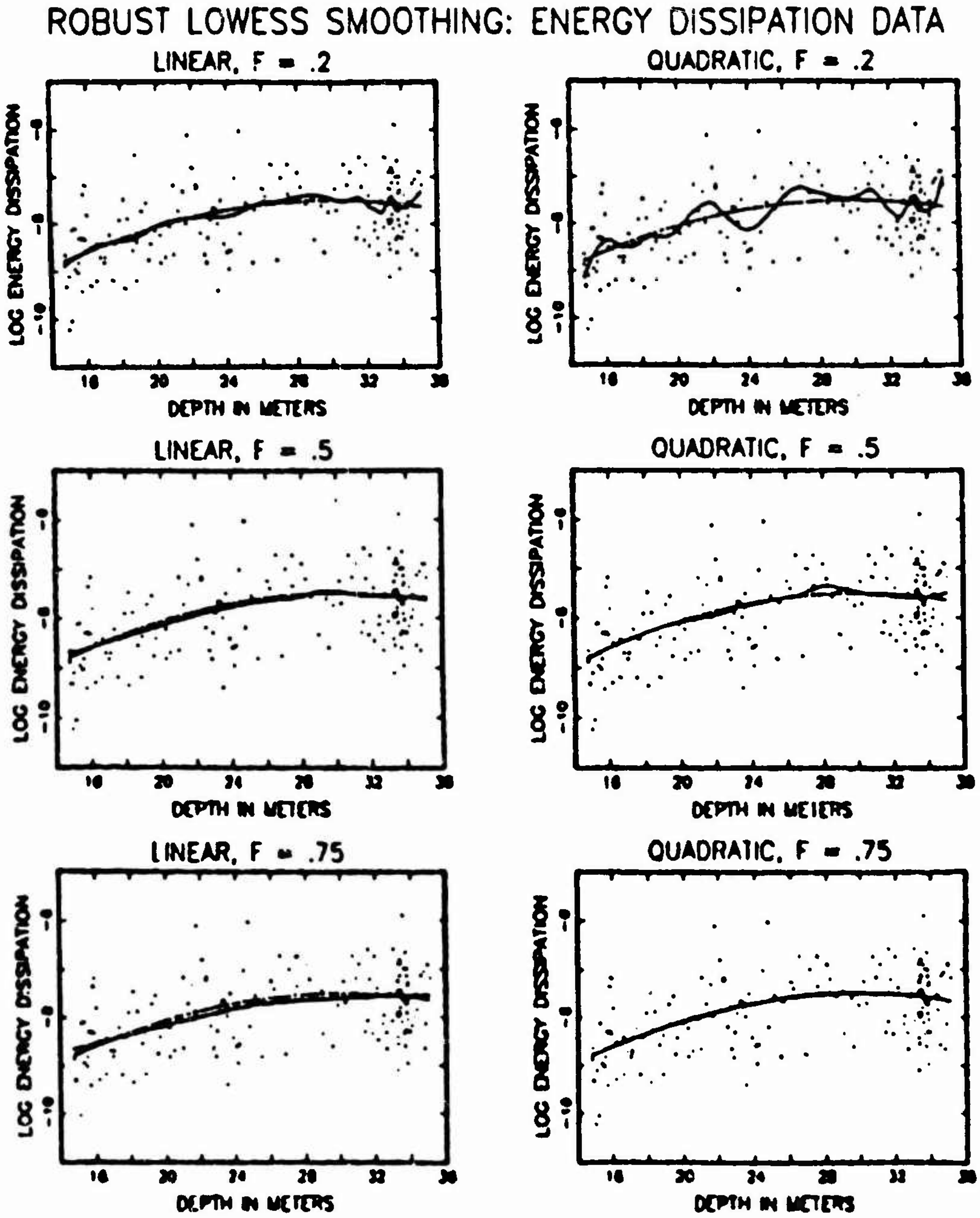


Figure 3.14 LOWESS Smoothing of Energy Dissipation Data Using Linear and Quadratic Regressions in Step Three.

The second irregularly shaped plot to be smoothed, a lag-1 plot of 200 NEAR(1) random variables, is pictured in Figure 3.16

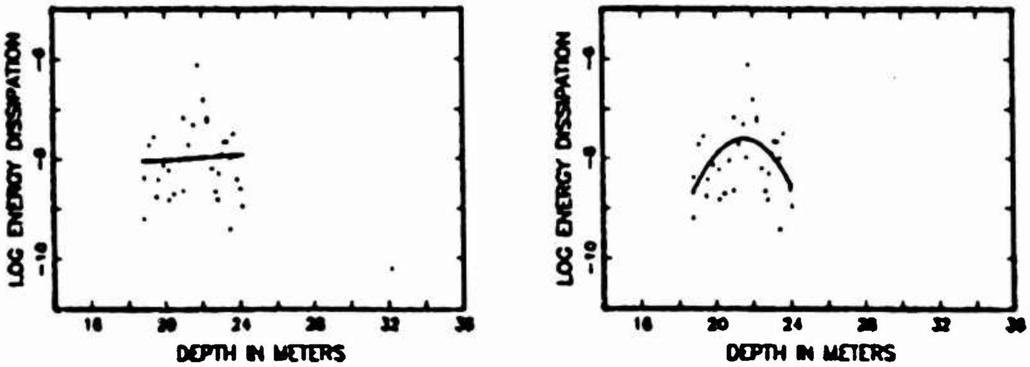


Figure 3.15 LOWESS Smoothing of X53 in Energy Dissipation Data Using Linear and Quadratic Regressions in Step Three.

TABLE III

Smirnov Test Comparing the Distribution of Residuals from Smoothing and Regression of Energy Data

type	F	T	Ks(.95)	.
lin	.50	.216	.149	reject
lin	.75	.156	.149	reject
quad	.50	.156	.149	reject
quad	.75	.078	.149	accept

The NEAR(1) process, derived by Lawrence and Lewis [Ref. 13], is a new first order autoregressive time series model with exponentially distributed marginals. NEAR(1) data is generated as a simple linear combination of a series, En, of independent exponential random variables by the model

$$X_N = \begin{cases} \epsilon_N + BX_{N-1} & \text{W.P. } A \\ 0 & \text{W.P. } (1-A) \end{cases} \qquad N = 0,1,2 \ldots$$

$$\epsilon_N = \begin{cases} E_N & \text{W.P. } \dfrac{1-B}{1-(1-A)B} \qquad N = 0,1,2 \ldots \\ (1-A)BE_N & \text{W.P. } \dfrac{AB}{1-(1-A)B} \end{cases}$$

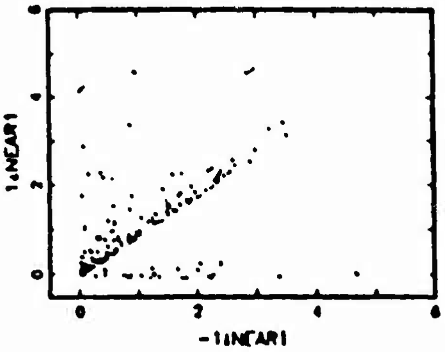


Figure 3.16 Lag-1 Plot of NEAR(1) Random Variables Having Autocorrelation .75.

These NEAR(1) variables have some interesting properties that make them especially suitable for testing smoothing routines. They have fixed serial lag-1 correlation, ρ_1 = AB and have conditional expectation

$$E[X_N | X_{N-1} = X] = (1-AB)\lambda^{-1} + ABX$$

The follcwing parameters were used to generate the variables for the test; A=.83, B=.9, λ= 1. A successful smooth of Figure 3.16 should prcduce a straight line of the form

$$Y = .25 + .75X$$

not at all what one would expect from looking at the plct.

Figure 3.17 presents comparison plots of robust and non-robust linear regression and robust and non-robust LOWESS smoothing of the near(1) data of Figure 3.16. The robust regression function contained in the IBM GRAFSTAT package was used in this example.

Examination of the plots in Figure 3.17 shows, once again, that LOWESS smooths are comparable to those produced by accepted linear regression techniques. It also reveals that neither the linear regression nor LOWESS procedures were able to reproduce the true lag-1 relationship, (Y = .25 + .75X), shown in the lower right hand plot. Both rcbust curves do present an accurate picture of where most of the data points lie, and could be used to predict where a majority of the future points are likely to fall. Relying on these curves, however, would probably lead to the conclusion that the points abcve and below these lines represent outliers, which may cr may not be the case.

It must be concluded from LOWESS' performance on these two data sets, however, that it smooths unequally spaced data as well as currently available regression techniques.

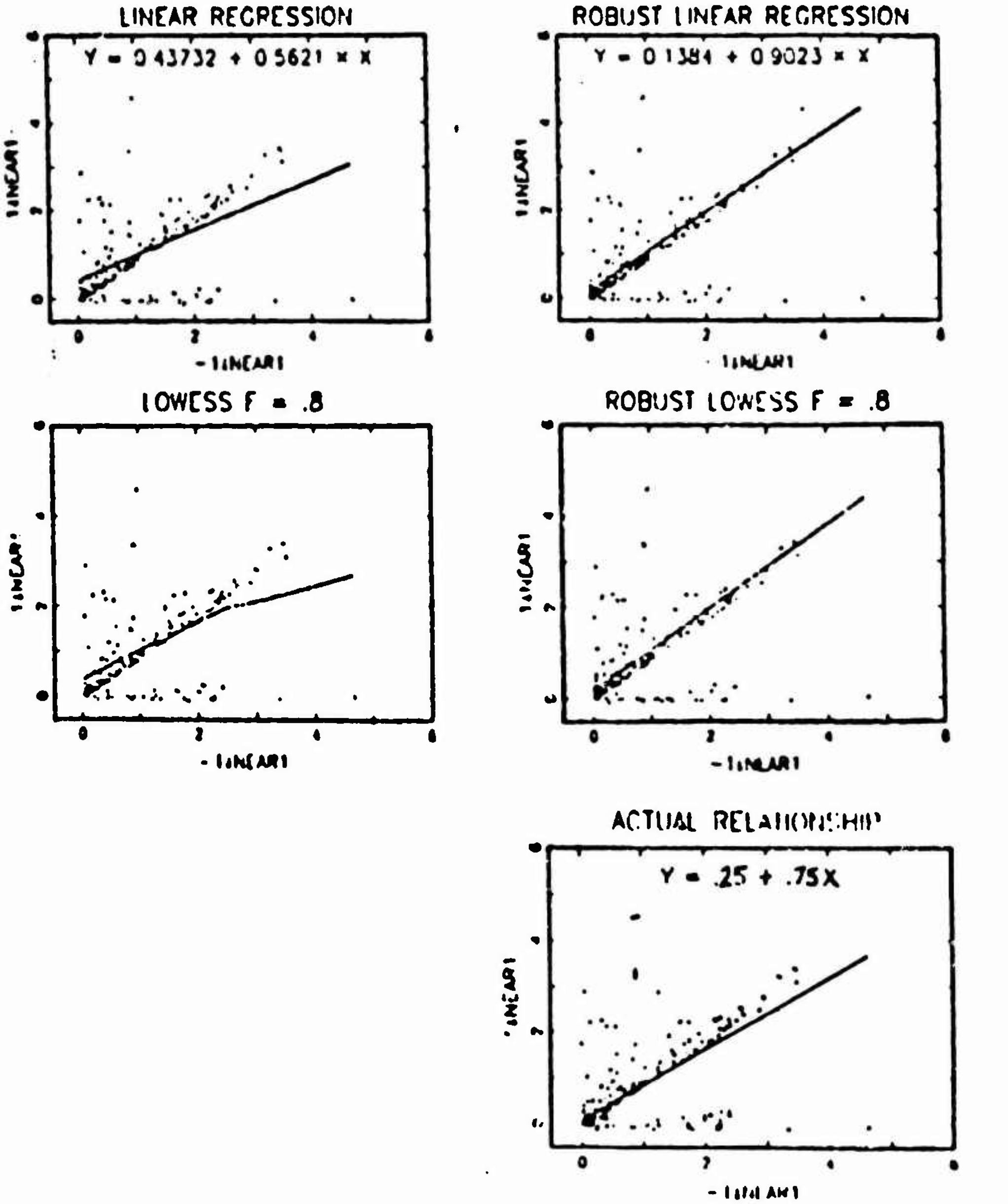


Figure 3.17 Comparison of Robust and Non-Robust Linear Regression and LOWESS Smoothing of the Lag-1 Plot of NEAR(1) Data.

IV. USING THE APL VERSION OF LOWESS

A. OVERVIEW

This chapter provides prospective users with detailed instructions for using LOWESS as a stand-alone program or in combination with the experimental GRAFSTAT graphics package. In either mode, LOWESS will provide the user with vectors of robust or non-robust smoothed $\hat{Y}i$ values and their associated residuals. When used in conjunction with GRAFSTAT, it will also produce a scatter plot of the original data with the LOWESS smoothed curve superimposed. A similar type presentation of the absolute value of the residuals versus Xi is also available on request from the program, Figure 4.1

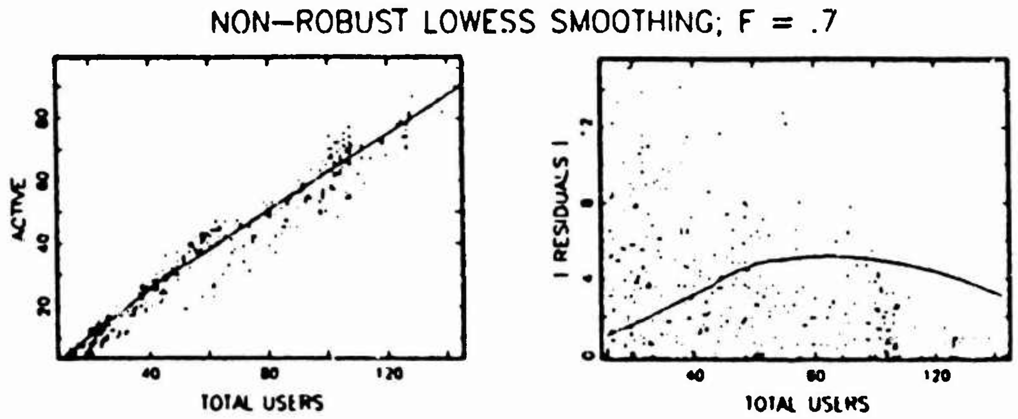


Figure 4.1 Sample of Graphical Outputs from LOWESS: Smooths of the Data (left), and Residuals (right).

LOWESS is a completely interactive program. All user defined parameters and option selections are entered in response to program queries. The stand-alone and combined graphics modes of operation are differentiated only by their initial set up procedures and by the choice of terminals on which the program is run.

Although no APL programming skills are required to operate LOWESS, users should become familiar with system commands and procedures for entering the APL environment, loading and copying workspaces and variables and for saving workspaces by reading appropriate sections of [Ref. 14]. Operating instructions presented in the follow-on sections of this chapter have been written for users who have had little or no experience with APL. Experienced users may find it more convient to refer to the summarized procedures presented in the Tables at the end of this chapter.

LOWESS is not a W.R Church computer center supported program and is not included in any of the APL libraries listed in [Ref. 15]. Interested users should contact Professor P.A.W. Lewis, Department of Operations Research, U.S. Naval Postgraduate School, for information concerning access to the APL workspace DTNLFNS. This workspace, which contains LOWESS and several other data analysis related programs, should be copied and stored on the user's A disk.

B. TERMINAL REQUIREMENTS

LOWESS, in the stand-alone mode can be run on any APL capable terminal at the U. S. Naval Postgraduate School. The IBM GRAFSTAT software, which generates the graphical displays when operating LOWESS in the combined graphics mode, requires the use of either IBM 3277GA or 3278/79 graphics display terminals. The 3278 terminals require special modification to produce graphical displays. None of these terminals are available for public use at the Naval Postgraduate School. See Table IV for a summary.

C. PROGRAM INITIALIZATION: STAND-ALONE MODE

Since LOWESS is written in APL, users must enter the APL sub-environment after completing normal log on procedures.

This is done by typing the letters "APL" and depressing the enter key. The response "CLEAR WS" indicates that the computer is ready to accept APL commands.

APL uses a special character set that is invoked by keying the APL ON/OFF key while depressing the ALT key on IBM 3278/79 terminals or by merely hitting the APL ON/OFF key on the 3277GA graphics display terminals. These special APL characters are imprinted in red (3278/79 terminals) or black (3277GA terminals) on the top and front surfaces of the normal keys. The symbols located on the front of the keys are accessed by typing the appropiate key while depressing the APL ALT key. When two APL characters are pictured on the top surface of the same key, the uppermost character is invoked by hitting that key while depressing the SHIFT key, much the same as producing capital letters during normal typing operations.

The final step in the initialization procedure consists of loading LOWESS and associated sub-programs into the active APL workspace. This is accomplished by entering the system command ")PCOPY DTNLFNS LOWESS "[1]. This command copies a group of programs required to execute LOWESS. See [Ref. 16 ,p.107], for information about the APL GROUP command. The computer responds by presenting WS size and "date-saved" information when all programs have been loaded. Initialization is now complete and the user is ready to execute LOWESS by typing "LOWESS" and hitting enter. From this point on, user enteries are made in response to program queries or instructions. Table I summarizes these initilization procedures.

[1] Underscored letters are obtained by typing the desired letter while depressing the APL ALT key.

D. PROGRAM INITIALIZATION: COMBINED GRAPHICS MODE

As noted in Section B of this chapter, the combined LOWESS-GRAFSTAT package can only be run on IBM 3277GA, 3279 or specially configured 3278 graphics display terminals. Additionally, efficient operation of GRAFSTAT requires a minimum workspace size of 2 megabytes. The W.R. Church Computer Center has established a limited number of public domain workspaces with special account numbers and passwords to meet this need, [Ref. 5]. Hard copy graphics printers are available for use with the 3277GA terminals located in Ingersall, Root and Spanegall Halls. The remainder ot this section focuses on the use of the 3277GA terminals.

Data files stored on the user's personal disk are unavailable for use while operating in one of the public workspaces. Users may:

1. send files to the public workspace's user number prior to logging on and commencing a work session;
2. link to his/her own disk after logging on to the public workspace useing CP link procedures outlined in [Ref. 17].

After logging on to one of the public workspaces and completing the data transfer or linking procedures described above, the user must enter the APL sub-environment by typing "APLGS7"[1] and hitting the enter key. The response, "CLEAR WS" indicates that the computer is ready to accept APL commands.

The special APL characters, labelled in black, are invoked by depressing the APL ON/OFF key. Since this key also turns the APL characters off, it may be necessary to check their status by trial and error. Detailed instructions

[1] The command, "APLGS7", invokes special system routines required to support the IBM GRAFSTAT software package. This procedure may change. Contact Professor P.A.W. Lewis, Department of Operations Research, if these procedures do not work.

for using the APL character set are presented in Section C of this chapter.

The initialization procedure is completed by loading GRAFSTAT and LOWESS into the active APL workspace. GRAFSTAT should be loaded first, by entering the system command ")LOAD GRAFSTAT". The GRAFSTAT package is quite large and may take several minutes to load. The following set of user instructions will appear on the screen when GRAFSTAT is fully loaded:

```
THIS IS A NEW (5/1/84)  RELEASE OF GRAFSTAT.  IT RUNS ON THE
3277/GA OR ON THE 3278/79. IT HAS A NUMBER OF NEW FUNCTIONS.
YOUR OLD CONTROL VECTORS WILL WORK  AS BEFORE.  IF YOU )COPY
RATHER  THAN  )LOAD  THIS WORKSPACE  YOU  MUST  EXECUTE  THE
FUNCTION LATENT BEFORE STARTING.  THE  NEXT  RELEASE  IS
SCHEDULED FOR 7/84.

TO BEGIN, TYPE: START

FOR MORE INFORMATION, TYPE: DESCRIBE
```

It is not necessary for the user to start, or even interact with GRAFSTAT to smooth a set of data: the GRAFSTAT message may be cleared by depressing the CLEAR key.

Users who have the APL workspace DTNLFNS stored on the public workspace disk, or who are linked to their own personal disk where it is stored, need only enter ")PCOPY DTNLFNS LOWESS " to complete the initialization process. The computer responds by presenting WS size and date saved information when all programs have been loaded. Initialization is now complete and the user is ready to execute LOWESS by typing "LOWESS" and hitting enter. From this point on user enteries are made in response to program queries or instructions. See Table VI for a summary of these procedures.

E. OPERATION OF LOWESS

This section provides detailed descriptions of the user inputs required during normal operation of LOWESS. The discussion assumes that one of the initialization procedures described in Sections C and D of this chapter has already been completed.

Execution of the LOWESS program is initiated by typing "LOWESS" and hitting the return key. Since the program is interactive it will respond with a series of queries or instructions requesting the user to input data or make decisions about the operation of the program. The exact sequence of program initiated queries and instructions is formulated in response to user inputs.

User-computer interactions required during execution of LOWESS are categorized into two types; data input and program operation.

Since the program cannot operate without data, the initial concern of LOWESS is to locate and read the data set it is about to smooth. Data can be read from the active APL workspace, a stored APL workspace or from a stored CMS file. Data that is not located in the active workspace must be accessible from that workspace. This presents no problem when the user is operating under his/her personal user number and the data is stored on his/her disk. This may become a problem when the user is logged on to one of the public workspaces described in Section D of this cahapter, and has not:

1. sent the data to the public workspace where he/she is working and stored it on the assoceated A disk;
2. linked to his/her own disk prior to entering the APL sub-environment, see Section D of this chapter.

Wherever the data is stored, it MUST be formatted into two separate lists, one containing the X values and the other containing the corresponding Y values of the points being smcothed.

Data which resides in the active workspace as APL vectors[1] is entered into LOWESS when the user types the variable name and hits enter in response to appropriate program requests.

Data which is stored in another APL workspace on the disk in use or on a disk to which the user is linked, will be transferred to the active workspace by the sub-program DATAINPUT. The user needs only to enter the workspace name and variable names when requested. DATAINPUT will also read and convert CMS files stored on the disk in use or on a disk to which the user is linked, provided they are formatted as described above and contain only numerical data. A mixture of alphabetic and numeric characters in a CMS data file will create an error and terminate execution of LOWESS. These data transfer features will work equally well in either mode of operation. The IBM GRAFSTAT program contains functions entitled CMS READ and CMS WRITE that will convert data in both directions when operating in the combined graphics mode. Users will generally not need to use this feature of GRAFSTAT, however.

Program operation inputs include:

1. the value of the parameter F (selection considerations are discussed in Chapter II Section C);
2. whether robust or non-robust smoothing is desired;
3. whether or nct a plot of the original data and smoothed curve is desired;

[1] In APL, a list of data points stored under a single variable name is referred to as a vector. See [Ref. 14], for further details.

4. whether or not a plot of the absolute values of the residuals and associated smoothed curve is desired;

5. X and Y axis labels for these plots.

Plots can only be generated while operating LOWESS in the combined graphics mode. Requesting plots when GRAFSTAT has not been loaded will produce an error and terminate execution. Hard copies of plots may be obtained by depressing the HARD COPY button on the bottom of the graphics screen.

TABLE IV

Summary of Terminal Requirements and Available Outputs

	Stand-Alone Mode	Combined Graphics
Terminal Required	3277GA 3278 3279	3277GA, 3279 or 3278 with graphics board
Additional Software Required	ncne	IBM GRAFSTAT pgm.
Available Output	Numerical: YSMTH .. smooth Y X1 ... original X Y1 ... original Y RESY .. residuals	Numerical: YSMTH .. smooth Y X1 ... original X Y1 ... original Y RESY .. residuals Graphical: Smooth curve |Residuals| vs Xi

TABLE V

Initializaticn Procedures, Stand-Alone Mode

Objective	User Inputs	Program Response
(1) enter APL environment	"APL"	"CLEAR WS"
(2) invoke APL characters	APL ON/OFF key	none
(3) load LOWESS and assoc. programs	)PCOPY DTNLFNS LOWESS	"saved (date) (time)"

TABLE VI

Initialization Procedures, Combined Graphics

Objective	User Inputs	Program Response
(1) enter APL environment	"APLGS7"	"CLEAR WS"
(2) invoke APL characters	APL ON/OFF key	none
(3) load GRAFSTAT	")LOAD GRAFSTAT"	initializaticn screen, see p 59
(4) load LOWESS	")PCOPY DTNLFNS LOWESS"	"saved (time) (date)"
(5) execute	"LOWESS"	

V. USING THE FORTRAN VERSION OF LOWESS

A. OVERVIEW

This chapter provides prospective users with detailed instructions for using a FORTRAN program that accomplishes the LOWESS curve smoothing procedure described in Chapter II. The program, entitled LOWESS, will provide the user with CMS files containing robust or non-robust Yi values and their associated residuals. These data files can be used to create plots of the raw and smoothed data points using DISPLA [Ref. 7], EASYPLOT, or other W.R. Church computer center supported IMSL or NON-IMSL plotting routines.

LOWESS is a completely interactive program. All user defined parameters and option selections are entered in response to program queries.

Although no FORTRAN programming skills are required to operate LOWESS, users should become familiar with FORTRAN and WATFIV operating system commands and also with the basic XEDIT editor, by reading appropriate sections of [Ref. 18], and [Ref. 19]. A limited ability to format, XEDIT and manipulate data files will be helpful when using LOWESS or when interacting with any of the plotting routines mentioned earlier.

B. TERMINAL REQUIREMENTS

LOWESS can be run on any remote terminal attached to the IBM computer located at the Naval Postgraduate School. The DISPLA and EASYPLOT plotting routines require the use of the IBM 3277GA graphics display terminals located in Ingersall, Root and Spanegall Halls. Plotting routines that use the remote VERSETEC or line printers can be accessed from any terminal.

C. PROGRAM INITIALIZATION (FORTRAN VERSION)

Since LOWESS is not a W.R. Church computer center supported program, it is not available in any of the center's public access libraries. Interested users should contact Professor P.A.W. Lewis, Department of Operations Research, U.S. Naval Postgraduate School, for information concerning access to LOWESS and its supporting programs. Copies of the programs listed in Table VII should be obtained and stored on the user's A disk. Annotated copies of the source codes are contained in Appendix (B).

TABLE VII

Programs and Subroutines Required for the Operation and Support of the FORTRAN Version of LOWESS

Filename	Filetype	Filemode
LOWESS	FORTRAN	A1
LOWS	EXEC	A1
PXSORT	FORTRAN	A1
LLBQF	FROTRAN	A1

PXSORT and LLBQF are contained in the IMSL library. Users having access to these programs through the W.R. Church computer center need not obtain personal copies.

The LOWS EXEC is used to activate system libraries, designate CMS storage space required for LOWESS input and output files. It is invoked by typing "LOWS EXEC" and hitting the ENTER key. The file definitions contained in the LOWS EXEC are listed in Table VIII. See [Ref. 17], for infomation on the use of EXEC executive programs.

This EXEC defines enough file space to accomodate five data sets. The user need only enter the appropriate file number when queried by LOWESS, to smooth any of the data sets.

TABLE VIII

Input and Output File Definitions Used in LOWS

File number	Filename	Filetype
2	LOW2	DATA
3	LOW3	DATA
4	LOW4	DATA
7	LOW7	DATA
8	LOW8	DATA

It may become necessary to change these filenames to avoid losing data when smoothing a large number of data sets or when smoothing one set a number of times. This may be accomplished in one of the following ways:

1. by entering the CMS ccamand "XEDIT LOWS EXEC" and changing the appropriate names;
2. by using the CMS command "R (old filename) (old filetype) (old filemode) (new filename) (new filetype) (new filemode)" for each file needing to be changed, see [Ref. 18].

File management is important. It is absolutely imperative that data input files have the same filename, filetype and filemode listed in the LOWS EXEC to prevent inadvertant smoothing of the wrong data or to prevent programming error.

D. DATA FILES (FORTRAN VERSION)

LOWESS requires that data be input in two columns of floating point constants in (2F15.5) format, X values on the left and Y values on the right. This is accomplished by creating a new file with the command "XEDIT (filename) (filetype)." The filename and filetype chosen should be one of those listed in Table VIII or one that is contained in the user's own LOWS EXEC. Refer to [Ref. 19], chapter 2, for more detailed instruction on creating files. The (2F15.5) format requires that all input variables contain a decimal point followed by no more than five decimal places. The X values must be entered in the first fifteen spaces and the Y values in the second fifteen spaces of each line (one set per line).

The output from LOWESS is placed in a file designated by the user. This can be the same file used for inputting the (X,Y) values or a different one. A different file should be used if the same data set is going to be smoothed with several different parameters. This output is printed in (4F15.3) format. The first column is the original X values ordered from smallest to largest. Column two contains the corresponding Y values, while column three contains the smoothed Yi values and column four contains the (Yi-Yi) residuals.

E. OPERATION OF LOWESS (FORTRAN VERSION)

This section provides detailed descriptions of the user inputs required during normal operation of LOWESS. The discussion assumes that the LOWS EXEC has been properly prepared and executed and that input files have been built according to instructions presented in Section C of this chapter.

Execution of the LOWESS program is initiated by typing "WATFIV LOWESS * (XTYPE". Since the program is interactive, it will respond with a series of queries or instructions requesting the user to input data or make decisions about the operation of the program.

The initial concern of LOWESS is to locate and read the data set it is about to smooth. Data can only be read from one of the files defined in the LOWS EXEC routine. The user tells LOWESS what file to read by entering the appropriate file number (2,3,4,7 or 8) in response to the instruction "ENTER THE FILE NUMBER OF THE INPUT DATA FILE." The program will terminate with an error if the LOWS EXEC was not properly prepared or if the data file was not formatted as describLed in the preceding section. Other program requested inputs include:

1. the value of the parameter F (selection considerations are discussed in Chapter II Section C);
2. whether or robust or non-robust smoothing is desired;
3. the file number of the desired output file.

APPENDIX A

APL PROGRAMS

This Appendix contains annotated listings of the APL programs written for this thesis. Source listings of the system library programs used to support the CMSREAD function called in the program DATAINPUT are not included.

LOWESS is an interactive program that executes the Robust-Locally-Weighted Regression Scatter-Plot Smoothing procedure described in the preceeding sections of this paper. It calls the following subprograms; DATAINPUT, REPEATCK, REGRES, REGRES2 PLOTQUERY and LOWS during execution. Refer to Chapter IV for detailed user instructions.

```
  ∇LOWESS
   [0]    LOWESS;N;Q;WX;J;I;A;B;Q;STRP;U;D;TX;WT;Z;BR;DA;DB;R;U1;M;RO;
                  AR;RHS;PROCEED;N1;PT;SKP;YS;F;ROB;REG;XAXIS;YAXIS;
                  PHDR;QS5;QS6;PT
   [1]   ⍝⍝⍝ DO NOT MOVE OR ERASE; GRAFSTAT FUNCTION HEADER
   [2]   ⍝⍝⍝ GRAFSTAT WILL NOT ADD A LINE TO THIS FUNCTION WITHOUT
   [3]   ⍝⍝⍝ THIS HEADER
   [4]   ⍝⍝⍝
   [5]   ⍝⍝⍝ LOWESS CALLS THE FOLLOWING PROGRAMS AND VARIABLES:
   [6]   ⍝⍝⍝ DATAINPUT; REPEATCK; PLOTQUERY; REGRES; REGRES2; RPLT;
   [7]   ⍝⍝⍝ NRPLT; RESPLT; SRESPLT
   [8]   ⍝⍝⍝
   [9]    ⎕PP←6
   [10]   DATAINPUT
  →[11]   →L9×ι(PROCEED≠'N')
  →[12]   →0
   [13]  L9:Y1←Y←Y[⍋X] ]  ORDER DATA
   [14]   X1←X←X[⍋X]  ]
   [15]   'INPUT F ... (0≤F≤1)'
   [16]   Q←⌊0.5+Q←(N1←⍴X)×F←⎕
   [17]   'DO YOU WANT TO USE LINEAR OR QUADRATIC FITTING DURING '
   [18]   'THIS SMOOTHING ROUTINE?'
   [19]   '(LIN OR QUAD)'
   [20]   REG←1↑⎕
   [21]   'DO YOU WANT TO USE THE ROBUST SMOOTHING OPTION?'
   [22]   '(YES OR NO)'
   [23]   ROB←1↑⎕
   [24]   YS←N1⍴0
   [25]   WX←N1⍴1
   [26]   J←0        ]  COUNTER FOR ROBUST SMOOTHING LOOP
   [27]  L1:J←J+1    ]
   [28]   I←0        ]
   [29]   A←1        ]  STARTS FIRST STRIP AT X1 ... XQ
   [30]   B←Q        ]
```

```
 [31]  L2:I←I+1    INCREMEMENTS THROUGH X_I ... X_N
→[32]   →L6×⍳(I>N1)
 [33]   REPEATCK  PREVENTS COMPUTATIONS OF Ŷ_I FOR REPEAT X_I
→[34]   →L5×⍳(SKP='Y')
 [35]   STRP←(A+(0,⍳(B-A)))
→[36]   →L3×⍳0≠D←⌈/|U←(X[I]∘.-X[STRP])   . COMPUTES D_I
 [37]   YS[I]←(+/(LST/Y))÷(+/LST←X=X[I])  USES AVG Y_I IF D_I=0
→[38]   →L5
 [39]  L3:WT←WX[STRP]×TX←((1-(|U*3))*3)×((|U←U÷D)<1) TRICUBE WT FCN
→[40]  L4:→R2×⍳(REG≠'L')               ┐
 [41]   X[STRP] REGRES Y[STRP]          │ WEIGHTED REGRESSIONS
→[42]   →L5                             │
 [43]  R2:X[STRP] REGRES2 Y[STRP]      ┘
→[44]  L5:→L2×⍳(B≥N1)∨(I≥N1)                            ┐
→[45]   →L2×⍳((DA←(X[I+1]-X[A]))≤(DB←(X[B+1]-X[I+1])))  │ADVANCE STRIP
 [46]   A←A+1                                           │
 [47]   B←B+1                                           ┘
→[48]   →L5
 [49]  L6:RO←|R[⍋(|R←RESY←(Y-YS))]               ┐
→[50]   →L10×⍳(0≠M←0.5×+/|(RO[(⌈N1÷2),1+⌊N1÷2]))  │
 [51]   U1←1                                      │ BICUBE  WT FCN
→[52]   →L11                                      │
 [53]  L10:U1←R÷(6×M)                             │
 [54]  L11:WX←((1-(U1*2))*2)×((|U1)<1)            ┘
→[55]   →L7×⍳(ROB≠'Y')
→[56]   →L1×⍳(J≤2)
 [57]  L7:PLOTQUERY RUN PLOTS
 [58]   YSMTH←YS
→[59]  ⍝→L8×⍳(PT≠'Y')
→[60]  ⍝→0
 [61]  L8:'THE OUTPUT FROM THIS LOWESS SMOOTHING IS STORED UNDER THE'
 [62]   'FOLLOWING VARIABLE NAMES:'
 [63]   '     YSMTH ....... SMOOTHED Y VALUES'
 [64]   '     X1 .......... X VALUES ARRANGED IN ASCENDING ORDER'
 [65]   '     Y1 .......... ORIGINAL Y VALUES'
 [66]   '     RESY ........ RESIDUALS'
```

DATAINPUT controls the data entry portion of the procedure. Data and program operating parameters are entered in response to program queries. DATAINPUT accepts data that is stored in the active APL workspace, transfers data from other APL workspaces and converts CMS data into APL.

```
**DATAINPUT
   [0]     DATAINPUT;QS1;QS2;QS3
   [1]     PROCEED←'Y'
   [2]     ' '
   [3]     'IS YOUR DATA SET LOCATED IN THIS WORKSPACE?'
   [4]     '(YES OR NO)'
   [5]     QS1←1↑⎕
  →[6]     →LP1×⍳(QS1='N')
   [7]     'ENTER THE NAME OF THE X VARIABLE'
   [8]     X←⎕
   [9]     'ENTER THE NAME OF THE Y VARIABLE'
   [10]    Y←⎕
  →[11]    →END
   [12]  LP1:'IS YOUR DATA LOCATED:'
   [13]   '    (1) IN AN APL WORKSPACE LOCATED ON THIS DISK OR ON A DISK'
   [14]   '        THAT YOU ARE LINKED TO;'
   [15]   '    (2) IN A CMS FILE ON THIS DISK OR ON A DISK THAT YOU ARE'
   [16]   '        LINKED TO;'
   [17]   '    (3) NEITHER (1) OR (2) ABOVE.'
   [18]   'ENTER (1,2 OR 3)'
   [19]   QS2←⎕
  →[20]   →(LP2,LP3,LP4)[QS2]
   [21]  LP2:'TO TRANSFER YOUR DATA TO THIS WORKSPACE:'
   [22]   '    (1) TYPE ...)PCOPY (WS NAME) (X VARIABLE NAME) (Y
      VARIABLE NAME)'
   [23]   '                EXAMPLE:  )PCOPY DATA  X  Y'
   [24]   '        IF YOUR DATA IS STORED AS TWO SEPERATE VARIABLES'
   [25]   '    (2) TYPE ...)PCOPY (WS NAME) (VARIABLE NAME)'
   [26]   '                EXAMPLE:  )PCOPY DATA ARRAY'
   [27]   '        IF YOUR DATA IS STORED UNDER A SINGLE VARIABLE NAME'
   [28]   '        AS IN A TWO DIMENSIONAL ARRAY'
   [29]   ' '
   [30]   '        DATE AND TIME SAVED INFORMATION IS DISPLAYED'
   [31]   '        WHEN THE TRANSFER IS COMPLETE. THEN ENTER    → GO
      '
   [32]   '        TO CONTINUE THE LOWESS SMOOTHING PROGRAM'
   [33]   S∆DATAINPUT←GO
   [34]  GO:'DO YOU NEED TO DEFINE YOUR X AND Y VARIABLES ANY FURTHER?'
   [35]   'ANSWER NO IF YOU ENTERED SEPARATE X AND Y VARIABLE NAMES'
   [36]   'IN THE PRECEDING STEP. OTHERWISE ANSWER YES.'
   [37]   '(YES OR NO)'
   [38]   QS3←1↑⎕
  →[39]   →END×⍳(QS3='N')
   [40]   'DEFINE THE X VARIABLE'
   [41]   X←⎕
   [42]   'DEFINE THE Y VARIABLE'
   [43]   Y←⎕
  →[44]   →END
   [45]  LP3:'TO TRANSFER YOUR CMS DATA FILE TO THIS WORKSPACE:'
   [46]   '    (1) ANSWER THE FOLLOWING QUESTIONS ABOUT YOUR X DATA FILE'
   [47]   X←CMSREAD
   [48]   '    (2) ANSWER THE FOLLOWING QUESTIONS ABOUT YOUR Y DATA FILE'
   [49]   Y←CMSREAD
   [50]   'YOU ARE NOW READY TO PROCEED WITH LOWESS'
  →[51]   →END
   [52]  LP4:'YOUR DATA MUST BE STORED IN AN APL WORKSPACE OR IN A CMS
      FILE'
   [53]   'LOCATED ON THIS DISK OR ON A DISK TO WHICH YOU ARE LINKED.
      LOWESS'
   [54]   'IS BEING TERMINATED. PLEASE COMPLY WITH CONDITION (1) OR (2)
      '
   [55]   'AND REINITIATE LOWESS.'
   [56]   PROCEED←'N'
   [57]  END:S∆DATAINPUT←0
```

REPEATCK reduces the number of computations required to smooth a data set by assigning the same smoothed Y value to data points that have the same X value.

```
·REPEATCK
 [0]   REPEATCK
 [1]   SKP←'N'
→[2]   →END×⍳(I≤1)
→[3]   →END×⍳(X[I]≠X[I-1])
 [4]   YS[I]←YS[I-1]
 [5]   SKP←'Y'
 [6]  END:
```

PLOTQUERY controls the the graphical output when operating with the IBM GRAFSTSAT statistical graphics package. It calls the sub program LOWS to smooth the absolute value of the (Yi-Yi) residuals obtained from smoothing the original data.

```
·**PLOTQUERY
 [0]    PLOTQUERY
 [1]    ' '
 [2]    'DO YOU WANT A PLOT OF YOUR LOWESS SMOOTHED CURVE?'
 [3]    '(YES OR NO) ..... ENTER NO IF NOT USING GRAFSTAT'
 [4]    PT←1↑⎕
→[5]    →END×⍳(PT≠'Y')
 [6]    'INPUT X AXIS LABEL'
 [7]    XAXIS←⎕
 [8]    'INPUT Y AXIS LABEL'
 [9]    YAXIS←⎕
→[10]   →PL1×⍳(ROB≠'Y')
 [11]   PHDR←'ROBUST LOWESS SMOOTHING; F = ',⍕F
 [12]   R̲UN RPLT
→[13]   →PL2
 [14]  PL1:PHDR←'NON-ROBUST LOWESS SMOOTHING; F = ',⍕F
 [15]   R̲UN NRPLT
 [16]  PL2:'DO YOU WANT A PLOT OF |RESIDUALS| VS X?'
 [17]   '(YES OR NO)'
 [18]   QS5←1↑⎕
→[19]   →END×⍳(QS5≠'Y')
 [20]   'DO YOU WANT THIS PLOT SMOOTHED?'
 [21]   '(YES OR NO)'
 [22]   QS6←1↑⎕
→[23]   →PL3×⍳(QS6≠'Y')
 [24]   X LOWS(|RESY)
 [25]   R̲UN SRESPLT
→[26]   →END
 [27]  PL3:R̲UN RESPLT
 [28]  END:
```

LOWS is used to smooth the $(Yi-\hat{Y}i)$ residuals obtained from smoothing the original data set. It operates exactly like LOWESS except for the data input and graphical output setctions.

```
 →→LOWS
   [0]     X LOWS Y;N1;Q;WX;J;I;A;B;Q;STRP;U;D;TX;WT;Z;BR;DA;DB;R;U1;M;
                    RO;AR;RHS;YZ
   [1]     Y←Y[⍋X]
   [2]     X←X[⍋X]
   [3]     Q←⌊0.5+Q←(N1←⍴X)×F
   [4]     YS←N1⍴0
   [5]     WX←N1⍴1
   [6]     J←0
   [7]    L1:J←J+1
   [8]     I←0
   [9]     A←1
   [10]    B←Q
   [11]   L2:I←I+1
  →[12]    →L6×ι(I>N1)
   [13]    REPEATCK
  →[14]    →L5×ι(SKP='Y')
   [15]    STRP←(A+(0,ι(B-A)))
  →[16]    →L3×ι0≠D←⌈/|U←(X[I]∘.-X[STRP])
   [17]    WT←WX[STRP]×TX←Q⍴1
   [18]    YS[I]←(+/(LST/Y)) ÷(+/LST←X=X[I+1]
  →[19]    →L5
   [20]   L3:WT←WX[STRP]×TX←((1-(|U*3))*3)×((|U←U÷D)<1)
  →[21]   L4:→R2×ι(REG≠'L')
   [22]    X[STRP] REGRES Y[STRP]
  →[23]    →L5
   [24]   R2:X[STRP] REGRES2 Y[STRP]
  →[25]   L5:→L2×ι(B≥N1)∨(I≥N1)
  →[26]    →L2×ι((DA←(X[I+1]-X[A]))≤(DB←(X[B+1]-X[I+1])))
   [27]    A←A+1
   [28]    B←B+1
  →[29]    →L5
   [30]   L6:RO←|R[⍋(|R←(Y-YS))]
  →[31]    →L10×ι(0≠M←0.5×+/|(RO[(⌈N1÷2),1+⌊N1÷2]))
   [32]    U1←1
  →[33]    →L11
   [34]   L10:U1←R÷(6×M)
   [35]   L11:WX←((1-(U1*2))*2)×((|U1)<1)
  →[36]    →L12×ι(ROB≠'Y')
  →[37]    →L1×ι(J≤2)
   [38]   L12:
```

REGRES computes linear least squares regressions of Y on X while REGRES2 computes quadratic least squares regressions of Y on X.

```
*REGRES
   [0]    XR REGRES YR;DEN;W1;B1;B2
   [1]    DEN←((+/W1)×(+/W1×XR*2))-((+/XR×W1←WT*0.5)*2)
  →[2]    →L1×⍳((|DEN)≥0.0001)
   [3]    YS[I]←(+/YR)÷ρYR
  →[4]    →0
   [5]   L1:B2←(((+/W1)×(+/(W1×XR×YR)))-((+/W1×XR)×(+/W1×YR)))÷DEN
   [6]    B1←((+/W1×YR)-B2×(+/W1×XR))÷(+/W1)
   [7]    YS[I]←B1+B2×X[I]
```

```
*REGRES2
   [0]    X2 REGRES2 Y2
   [1]    A1←(+/X2×(WT*0.5))
   [2]    A2←(+/(X2*2)×(WT*0.5))
   [3]    A3←(+/(X2*3)×(WT*0.5))
   [4]    AR2← 3 3 ρ(+/WT*0.5),A1,A2,A1,A2,A3,A2,A3,(+/(X2*4)×(WT*0.5))
   [5]    RHS2←(+/Y2×WT*0.5),(+/X2×Y2×WT*0.5)
   [6]    RHS2← 3 1 ρRHS2,(+/(X2*2)×Y2×WT*0.5)
   [7]    BR←RHS2⌹AR2
   [8]    YS[I]←BR[1;1]+(BR[2;1]×X[I])+(BR[3;1]×X[I]*2)
```

The following character strings are the screen vectors used by the RUN function of GRAFSTAT to produce the plots of the LCWESS smoothe curves of the original data and absolute value of the residuals.

```
**NRPLT               73     CHARACTER
  *4♥X1♥Y1;YS♥0 1♥1♥.*+×▽△○●♦♦♥''♥FHDR♥XAXIS♥YAXIS♥21♥LIN♥LIN♥1 1 1♥0 1
       0 0


**RESPLT              80     CHARACTER
  **1♥X♥(|RESY)♥0♥1♥.*+×▽△○●♦♦♥''♥''♥XAXIS♥'|RESIDUALS|'♥22♥LIN♥LIN♥1 1
       1♥0 1 0 0♥


**RPLT                73     CHARACTER
  *4♥X1♥Y1;YS♥0 1♥1♥.*+×▽△○●♦♦♥''♥PHDR♥XAXIS♥YAXIS♥21♥LIN♥LIN♥1 1 1♥0 1
       0 0


**SRESPLT             85     CHARACTER
  **1♥X♥(|RESY);YS♥0
       1♥1♥.*+×▽△○●♦♦♥''♥''♥XAXIS♥'|RESIDUALS|'♥22♥LIN♥LIN♥1 1 1♥0 1 0
       0♥
```

APPENDIX B

FORTRAN PROGRAMS

This appendix contains a listing of the FORTRAN program and subroutine written to support this thesis. IMSL programs, LLBQF and PXSORT, used to support the LOWESS program are not listed. Detailed user instructions for operating these programs are contained in Chapter V.

```
$JOB C
      REAL
X(200),Y(200),YS(200)/200*0.0/,WX(200)/200*1.0/,A(2,2),B(2,1)
      C,U(200)/200*0.0/,D,U1,TX(200)/200*0.0/,WT(200)/200*0.0/
      C,WK(22),DA,DB,R(200)/200*0.0/,R1(200)/200*0.0/,RU,F,C(4)
      C(,W,BETA  (2,1),RED C
      INTEGER
AX,BX,A1,Q,I1,I2,I3,I4,I5,I6,I7,I8,I9,I10,N,IWK(2),IER,ROB
      C,IF1,IF2 C
      DATA AX/1/,ROB/-1/,N/0/ C
      F=.33
      IF1=2
      IF2=4
      N=0
    1 N=N+1
      READ(IF1,901,END=2)X(N),Y(N)
      GO TO 1
    2 N=N-1
      CALL XYSORT(X,Y,1,N)
      Q=IFIX((FLOAT(N)*F)+.5)
    4 CONTINUE
      AX=1
      A1=(AX-1)
      BX=0
      DO 65 I1=1,N
         I2=0
         D=0.0
         DO 10 I3=AX,BX
            I2=I2+1
            U(I2)=X(I1)-X(I3)
            IF(.NOT.ABS(U(I2)).GE.D)GO TO 5
               D=ABS(U(I2))
    5       CONTINUE
   10    CONTINUE
         IF(.NOT.D.GT.0.00001)GO TO 30
            DO 25 I4=1,Q
               U1=ABS(U(I4)/D)
               IF(.NOT.U1.LT.1.0)GO TO 15
                  TX(I4)=(1.0-(U1**3))**3
                  WT(I4)=TX(I4)*WX(A1+I4)
                  GO TO 20
   15          CONTINUE
                  TX(I4)=0.0
                  WT(I4)=0.0
   20          CONTINUE
   25       CONTINUE
```

```
               GO TO 40
 30         CONTINUE
               DO 35 I5=1,Q
                  TX(I5)=1.0
                  WT(I5)=WX(A1+I5)
 35            CONTINUE
 40         CONTINUE C
            A(1,1)=0.0
            A(1,2)=0.0
            A(2,1)=0.0
            A(2,2)=0.0
            B(1,1)=0.0
            B(2,1)=0.0
            DO 45 I6=1,Q
               I7=A1+I6
               W=SQRT(WT(I6))
               A(1,1)=A(1,1)+W
               A(1,2)=A(1,2)+(X(I7)*W)
               A(2,2)=A(2,2)+(W*(X(I7)**2))
               B(1,1)=B(1,1)+(Y(I7)*W)
               B(2,1)=B(2,1)+(Y(I7)*X(I7)*W)
 45         CCNTINUE
            A(2,1)=A(1,2) C
            CALL LLBQF(A,2,2,2,B,2,1,0,C,BETA,2,IWK,WK,IER) C
            YS(I1)=BETA(1,1)+BETA(2,1)*X(I1)
 50            CONTINUE
            IF(BX.GE.N)GC TO 60
            IF(I1.GE.N)GC TO 60
               DA=X(I1+1)-X(AX)
               DB=X(BX+1)-X(I1+1)
               IF(.NOT.DA.GT.DB)GO TO 55
                  AX=AX+1
                  BX=BX+1
                  GO TO 50
 55            CONTINUE
 60         CONTINUE
            A1=(AX-1)
 65      CCNTINUE C
         DO 70 I8=1,N
            R(I8)=Y(I8)-YS(I8)
            R1(I8)=ABS(R(I8))
 70      CCNTINUE C
         CALL PXSORT(R1,1,N) C
         L1=(N+1)/2
         L2=(N+2)/2
         MED=(R1(L1)+R1(I2))/2.0
         DO 85 I9=1,N
           IF((R1(I9).GT.0.0).AND.(ABS(MED).GT.0.0))GO TO 71
               WX(I9)=1.0
               GO TO 80
 71         RU=R(I9)/(6.C*MED)
            IF(.NOT.ABS(RU).LT.1.0)GO TO 75
               WX(I9)=(1.0-(RU**2))**2
               GO TO 80
 75         CCNTINUE
               WX(I9)=0.0
 80         CCNTINUE C
 85      CONTINUE C TEST
         WRITE(6,991)(WX(L)L=1,N)
991      FORMAT(1X,10F7.3) C END TEST C
         FCB=ROB+1 C      IF(.NOT.ROB.GE.2)GO TO 4
         DO 90 I10=1,N
            WRITE(IF2,900)X(I10),Y(I10),YS(I10)
 90      CCNTINUE
         STCP
900      FORMAT(1X,3F15.3)
901      FCFMAT(2F15.3)
         ENC C
         SUBROUTINE XYSCFT(A,B,II,JJ) C
```

```
      DIMENSION A(JJ),B(JJ),IU(16),IL(16)
      M=1
      I=II
      J=JJ
    5 IF(I .GE. J)GO TO 70
   10 K=I
      IJ=(I+J)/2
      T=A(IJ)
      T1=B(IJ)
      IF(A(I) .LE. T) GO TO 20
      A(IJ)=A(I)
      B(IJ)=B(I)
      A(I)=T
      B(I)=T1
      T=A(IJ)
      T1=B(IJ)
   20 L=J
      IF(A(J) .GE. T) GO TO 40
      A(IJ)=A(J)
      B(IJ)=B(J)
      A(J)=T
      B(J)=T1
      T=A(IJ)
      T1=B(IJ)
      IF(A(I) .LE. T) GO TO 40
      A(IJ)=A(I)
      B(IJ)=B(I)
      A(I)=T
      B(I)=T1
      T=A(IJ)
      T1=B(IJ)
      GO TO 40
   30 TT=A(L)
      TT1=B(L)
      A(L)=A(K)
      B(L)=B(K)
      A(K)=TT
      B(K)=TT1
   40 L=L-1
      IF(A(L) .GT. T) GO TO 40
   50 K=K+1
      IF(A(K) .LT. T) GO TO 50
      IF(K .LE. L) GO TO 30
      IF(L-I .LE. J-K) GO TO 60
      IL(M)=I
      IU(M)=L
      I=K
      M=M+1
      GO TO 80
   60 IL(M)=K
      IU(M)=J
      J=L
      M=M+1
      GO TO 80
   70 M=M-1
      IF(M .EQ. 0) RETURN
      I=IL(M)
      J=IU(M)
   80 IF(J-I .GE. 11)GO TO 10
      IF(I .EQ. II) GC TO 5
      I=I-1
   90 I=I+1
      IF(I .EQ. J) GC TO 70
      IF(A(I) .LE. A(I+1)) GO TO 90
      T = A(I+1)
      T1=B(I+1)
      K=I
  100 A(K+1)=A(K)
      B(K+1)=B(K)
      K=K-1
      IF(T .LT. A(K)) GO TO 100
      A(K+1)=T
      B(K+1)=T1
      GO TO 90
      END $ENTRY
```

The following LCWS EXEC routine sets the file definitions and invokes the appropriate systems libraries required to execute LOWESS. This routine is executed by typing "LOWS EXEC."

```
      GLCBAL MACLIB IMSLSP NONIMSL
      FILEDEF 02 DISK LOW2 DATA A (PERM
      FILEDEF 03 DISK LOW3 DATA A (PERM
      FILEDEF 04 DISK LOW4 DATA A (PERM
      FILEDEF 07 DISK LOW7 DATA A (PERM
      FILEDEF 08 DISK LOW8 DATA A (PERM
```

APPENDIX C

DATA SETS

This appendix contains four data sets that were used to compare LOWESS with MOVING AVERAGE, COSINE ARCH and LEAST SQUARES REGRESSION rooutines in Chapter III. They include:

1. TEST SET ONE ... used to test LOWESS' ability to detect and follow linear trends.
2. TEST SET TWO ... used to check LOWESS' performance on data sets that contain abrupt changes in curvature.
3. TEST SET THREE ... used to test LOWESS' ability to follow smooth changes in curvature.
4. Lag-1 points from NEAR(1) data ... used to check LOWESS' performance on unequally spaced data.

TABLE IX

Data Set One

X	Y	X	Y	X	Y
.200	-.398	10.200	8.696	20.200	21.520
.400	-.811	10.400	10.305	20.400	19.996
.600	-.103	10.600	10.997	20.600	21.018
.800	1.156	10.800	10.273	20.800	21.047
1.000	1.653	11.000	11.345	21.000	21.704
1.200	1.416	11.200	10.477	21.200	21.832
1.400	1.136	11.400	12.668	21.400	20.408
1.600	3.402	11.600	11.569	21.600	23.367
1.800	1.157	11.800	12.578	21.800	21.418
2.000	2.110	12.000	14.180	22.000	21.089
2.200	1.481	12.200	12.638	22.200	21.204
2.400	2.821	12.400	13.733	22.400	23.595
2.600	.669	12.600	12.851	22.600	22.441
2.800	3.460	12.800	12.490	22.800	25.504
3.000	1.897	13.000	12.077	23.000	22.802
3.200	3.097	13.200	12.815	23.200	23.059
3.400	2.340	13.400	14.558	23.400	23.811
3.600	2.361	13.600	14.463	23.600	22.421
3.800	1.911	13.800	12.765	23.800	23.522
4.000	3.026	14.000	13.807	24.000	22.419
4.200	4.412	14.200	12.900	24.200	25.249
4.400	4.893	14.400	14.707	24.400	24.703
4.600	6.147	14.600	15.569	24.600	23.373
4.800	5.445	14.800	14.053	24.800	24.870
5.000	2.852	15.000	12.204	25.000	24.603
5.200	4.171	15.200	15.897	25.200	26.589
5.400	5.258	15.400	18.607	25.400	26.764
5.600	3.073	15.600	16.136	25.600	26.258
5.800	5.487	15.800	16.098	25.800	26.291
6.000	5.406	16.000	16.284	26.000	26.801
6.200	6.532	16.200	17.160	26.200	25.433
6.400	6.959	16.400	18.488	26.400	26.764
6.600	7.500	16.600	18.125	26.600	26.202
6.800	6.599	16.800	16.605	26.800	27.664
7.000	6.766	17.000	17.017	27.000	26.822
7.200	8.650	17.200	17.446	27.200	29.074
7.400	9.236	17.400	16.546	27.400	27.572
7.600	7.217	17.600	18.758	27.600	28.872
7.800	7.955	17.800	17.962	27.800	27.765
8.000	7.035	18.000	19.557	28.000	26.499
8.200	8.239	18.200	18.006	28.200	28.565
8.400	9.165	18.400	20.051	28.400	28.201
8.600	8.005	18.600	16.701	28.600	27.210
8.800	8.930	18.800	20.623	28.800	29.029
9.000	9.035	19.000	17.482	29.000	29.271
9.200	8.575	19.200	18.149	29.200	28.834
9.400	8.860	19.400	19.450	29.400	30.777
9.600	11.480	19.600	18.145	29.600	28.802
9.800	8.796	19.800	20.267	29.800	28.863
10.000	9.503	20.000	20.545	30.000	29.998

TABLE X

Data Set Two

X	Y	X	Y	X	Y	X	Y
.200	¯.462	11.200	3.849	22.200	4.819	33.200	1.657
.400	¯2.191	11.400	4.554	22.400	4.469	33.400	2.245
.600	1.405	11.600	3.182	22.600	4.997	33.600	.862
.800	.947	11.800	3.159	22.800	6.256	33.800	3.226
1.000	.475	12.000	4.518	23.000	6.278	34.000	1.362
1.200	.832	12.200	5.736	23.200	6.490	34.200	2.923
1.400	¯.137	12.400	4.989	23.400	5.499	34.400	2.736
1.600	2.336	12.600	3.752	23.600	5.860	34.600	1.736
1.800	.779	12.800	5.165	23.800	4.325	34.800	2.129
2.000	2.597	13.000	4.052	24.000	4.949	35.000	1.433
2.200	1.144	13.200	3.594	24.200	6.690	35.200	1.313
2.400	1.832	13.400	3.895	24.400	6.339	35.400	2.756
2.600	¯.406	13.600	3.747	24.600	5.899	35.600	1.576
2.800	.419	13.800	4.171	24.800	4.233	35.800	.363
3.000	2.446	14.000	4.962	25.000	5.825	36.000	2.955
3.200	.641	14.200	3.356	25.200	5.742	36.200	.266
3.400	1.937	14.400	4.792	25.400	4.873	36.400	1.664
3.600	1.080	14.600	5.593	25.600	5.497	36.600	.323
3.800	1.384	14.800	4.630	25.800	7.697	36.800	.783
4.000	.251	15.000	5.203	26.000	4.600	37.000	1.419
4.200	.419	15.200	4.468	26.200	3.374	37.200	1.997
4.400	2.745	15.400	6.558	26.400	2.242	37.400	.533
4.600	1.795	15.600	5.484	26.600	4.078	37.600	1.137
4.800	1.121	15.800	2.766	26.800	4.090	37.800	.506
5.000	1.235	16.000	4.635	27.000	3.519	38.000	.671
5.200	2.942	16.200	2.812	27.200	6.651	38.200	¯.612
5.400	2.104	16.400	5.668	27.400	5.513	38.400	.376
5.600	2.753	16.600	5.055	27.600	5.141	38.600	1.921
5.800	2.717	16.800	5.319	27.800	4.818	38.800	¯.476
6.000	3.156	17.000	5.574	28.000	1.451	39.000	¯1.014
6.200	2.880	17.200	6.472	28.200	5.936	39.200	1.788
6.400	1.219	17.400	4.420	28.400	4.205	39.400	1.306
6.600	3.015	17.600	4.623	28.600	3.202	39.600	.853
6.800	3.845	17.800	5.396	28.800	1.977	39.800	¯1.468
7.000	3.529	18.000	5.778	29.000	4.046	40.000	1.554
7.200	.503	18.200	3.765	29.200	5.971	40.200	¯.542
7.400	2.686	18.400	4.290	29.400	4.175	40.400	¯2.351
7.600	2.717	18.600	4.900	29.600	4.583	40.600	1.165
7.800	3.438	18.800	2.397	29.800	3.479	40.800	.627
8.000	2.689	19.000	6.059	30.000	4.621	41.000	.075
8.200	3.278	19.200	3.894	30.200	1.989	41.200	.352
8.400	4.967	19.400	6.093	30.400	4.408	41.400	¯.697
8.600	4.288	19.600	4.174	30.600	3.896	41.600	1.696
8.800	3.788	19.800	5.615	30.800	3.112	41.800	.059
9.000	2.677	20.000	5.820	31.000	3.422	42.000	1.797
9.200	3.610	20.200	4.844	31.200	4.740	42.200	.264
9.400	3.908	20.400	5.602	31.400	3.108	42.400	.872
9.600	3.283	20.600	4.933	31.600	3.892	42.600	¯1.446
9.800	3.583	20.800	5.634	31.800	1.630	42.800	¯.701
10.000	4.415	21.000	4.003	32.000	4.039	43.000	1.246
10.200	5.578	21.200	4.389	32.200	4.600	43.200	¯.639
10.400	1.596	21.400	6.545	32.400	2.125	43.400	.577
10.600	2.962	21.600	4.546	32.600	1.625	43.600	¯.360
10.800	5.203	21.800	5.417	32.800	1.602	43.800	¯.136
11.000	4.682	22.000	3.613	33.000	3.180	44.000	¯1.349

TABLE XI

Data Set Three

X	Y	X	Y	X	Y
.063	.261	2.135	.560	4.208	¯1.733
.126	¯.129	2.198	.716	4.270	¯.860
.188	.053	2.261	1.376	4.333	.049
.251	¯.293	2.324	.410	4.396	¯.870
.314	1.316	2.386	.988	4.459	¯1.282
.377	1.340	2.449	.326	4.522	¯1.701
.440	¯.335	2.512	.875	4.584	¯1.025
.502	1.451	2.575	.175	4.647	¯.811
.565	.088	2.638	1.079	4.710	¯.891
.628	.435	2.700	.520	4.773	¯1.088
.691	.915	2.763	1.167	4.836	¯.980
.754	.522	2.826	.471	4.898	¯.662
.816	1.398	2.889	.684	4.961	¯.508
.879	1.381	2.952	.835	5.024	¯1.729
.942	.011	3.014	.344	5.087	¯.599
1.005	.310	3.077	¯.129	5.150	¯1.211
1.068	.496	3.140	¯.055	5.212	¯.595
1.130	1.115	3.203	¯.543	5.275	¯1.151
1.193	.713	3.266	¯1.152	5.338	¯.195
1.256	1.304	3.328	¯.111	5.401	¯.275
1.319	1.082	3.391	.024	5.464	¯1.133
1.382	.474	3.454	¯.180	5.526	¯.982
1.444	1.062	3.517	¯.520	5.589	.206
1.507	.624	3.580	¯.633	5.652	¯.113
1.570	.686	3.642	.088	5.715	¯1.503
1.633	1.695	3.705	¯.339	5.778	¯.228
1.696	.168	3.768	.216	5.840	¯.232
1.758	¯.025	3.831	¯.223	5.903	¯.824
1.821	1.215	3.894	.052	5.966	¯.949
1.884	.174	3.956	¯1.417	6.029	¯.078
1.947	.860	4.019	¯.899	6.092	¯.788
2.010	1.028	4.082	¯.310	6.154	.205
2.072	.743	4.145	.074	6.217	¯.100

TABLE XII

Lag-1 Data derived from NEAR(1) Process

X	Y	X	Y	X	Y	X	Y
1.020	.466	.871	.822	.563	.650	.313	.304
.035	1.020	.747	.871	.049	.563	.376	.313
.129	.035	1.385	.747	.133	.949	.329	.376
.125	.129	1.189	1.385	.334	.133	.363	.329
.153	.125	.017	1.189	.596	.334	.556	.363
.233	.153	.261	.017	.604	.596	.655	.556
2.077	.233	.366	.261	.527	.604	.544	.655
2.155	2.077	.349	.366	.934	.527	.569	.544
1.821	2.155	.364	.349	1.797	.934	.531	.569
.042	1.821	1.140	.364	1.496	1.797	.518	.531
.036	.042	1.020	1.140	1.420	1.496	.584	.518
.061	.036	3.508	1.020	1.522	1.420	4.292	.584
.149	.061	3.122	3.508	1.353	1.522	3.610	4.292
4.260	.149	2.623	3.122	1.187	1.353	4.074	3.610
4.095	4.260	2.654	2.623	1.050	1.187	3.492	4.074
3.422	4.095	.209	2.654	.898	1.050	3.644	3.492
2.854	3.422	.255	.209	.854	.898	3.147	3.644
2.609	2.854	.271	.255	1.631	.854	.022	3.147
2.176	2.609	1.185	.271	1.363	1.631	.330	.022
1.823	2.176	.989	1.185	1.172	1.363	.310	.330
1.617	1.823	2.867	.989	1.303	1.172	.597	.310
2.439	1.617	2.488	2.867	1.229	1.303	.551	.597
2.047	2.439	2.086	2.488	1.061	1.229	.544	.551
1.840	2.047	1.756	2.086	.962	1.061	.817	.544
3.049	1.840	1.530	1.756	.907	.962	.808	.817
2.682	3.049	1.456	1.530	.856	.907	.715	.808
2.239	2.682	.180	1.456	1.135	.856	.601	.715
1.889	2.239	.429	.180	.953	1.135	.618	.601
1.577	1.889	.031	.429	1.728	.953	1.525	.618
1.664	1.577	2.951	.031	.010	1.728	1.526	1.525
.103	1.664	2.565	2.951	.073	.010	1.279	1.526
.133	.103	2.133	2.565	.082	.073	1.065	1.279
.145	.133	3.737	2.133	.096	.082	.929	1.065
.207	.145	3.180	3.737	.098	.096	.814	.929
.221	.207	2.675	3.180	.234	.098	.703	.814
.196	.221	2.307	2.675	1.046	.234	.704	.703
.170	.196	1.996	2.307	1.017	1.046	.898	.704
.185	.170	1.892	1.996	1.239	1.017	.785	.898
.087	.185	1.700	1.892	.105	1.239	1.065	.785
2.258	.087	1.716	1.700	.124	.105	.995	1.065
1.938	2.258	1.599	1.716	.122	.124	3.157	.995
1.617	1.938	1.498	1.599	.122	.122	2.710	3.157
1.346	1.617	1.247	1.498	.154	.122	2.265	2.710
1.184	1.346	.044	1.247	.165	.154	1.883	2.265
1.007	1.184	.306	.044	.205	.165	1.566	1.883
.853	1.007	.255	.306	.190	.205	1.488	1.566
.779	.853	.258	.255	.315	.190	1.268	1.488
.727	.779	.519	.258	.335	.315	1.206	1.268
.822	.727	.650	.519	.304	.335	2.825	1.206

LIST OF REFERENCES

1. Chambers, J.M. and others, *Graphical Methods for Data Analysis*, Wadsworth, 1983.

2. Anscombe, F.J., "Graphs in Statistical Analysis," *The American Statistician*, volume 27, number 1, Feb. 1973.

3. Cleveland, W.S., "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, volume 74, number 368, Dec. 1979.

4. Kendall, M.G. and Stuart, A., *The Advanced Theory of Statistics*, volume 3, Hafner, 1966.

5. "Grafstat of the IBM 3277 Display," *Computer Center Newsletter*, Naval Postgraduate School, 18 May 1983.

6. Johnson, M.D. Jr., *Draftsman Displays, a Graphical Technique for Exploratory Data Analysis*, M.S. Thesis, Naval Postgraduate School, Monterey, Ca., June 1984.

7. Naval Postgraduate School, Technical Note VM-12, *Using Displa at NPS*, Jun. 1983.

8. Beaton, A.E. and Tukey, J.W., "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Teconometrics*, volume 16, 1974.

9. Rice, J., "Methods of Bandwidth Choice in Nonparametric Kernel Regression," *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, North-Holland, 1983.

10. Koopmans, L.H., *The Spectral Analysis of Time Series*, Academic Press, 1974.

11. Anscombe, F.J., *Computing in Statistical Science Through APL*, Springer-Verlag, 1981.

12. Conover, W.J., *Practical Nonparametric Statistics, Second Edition*, John Wiley and Sons, 1980.

13. Lawrence, A.J. and Lewis, P.A.W., "A New Autoregressive Time Series Model in Exponential Variables (NEAR(1))," *Advanced Applied Probability*, number 13, 1981.

14. Ramsey, J.B. and Musgrave, G.L., APL-STAT, Lifetime Learning Publications, 1981.

15. Naval Postgraduate School, Technical Note VM-10, VSAPL at NPS, July 1982.

16. APL Language, Fifth Edition, IBM, 1978.

17. Naval Postgraduate School, Technical Note TN-VM-04, Using Exec Files under CMS, Feb. 1984.

18. Naval Postgraduate School Technical Note VM-01 Users Guide to VM/CMS at NPS, Jul. 1984.

19. Naval Postgraduate School, Technical Note TN-VM-05, Introduction to the XEDIT Editor, Jul. 1983.

20. Naval Postgraduate School, Note MVS-02, Using the VERSATEC Plotter at NPS, Oct. 1983.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22314	2
2.	Superintendent Attn: Library Code 0142 Naval Postgraduate School Monterey, California 93943	2
3.	Professor P.A.W. Lewis Department of Operations Research Code 55Lw Naval Postgraduate School Monterey, California 93943	10
4.	Professor D. Gaver Department of Operations Research Code 55Gv Naval Postgraduate School Monterey, California 93943	1
5.	Professor R.R. Read Department of Operations Research Code 55Re Naval Postgraduate School Monterey, California 93943	1
6.	Associate Professor P.A. Jacobs Department of Operations Research Code 55Jc Naval Postgraduate School Monterey, California 93943	1
7.	Associate Professor R. Richards Department of Operations Research Code 55Rh Naval Postgraduate School Monterey, California 93943	1
8.	LCdr. P. Fishbeck, USN Department of Operations Research Code 55Fb Naval Postgraduate School Monterey, California 93943	1
9.	Cdr. G.W. Moran, USN 26 Martel Road Springfield, Massachusetts 01119	4